DmB 1.0

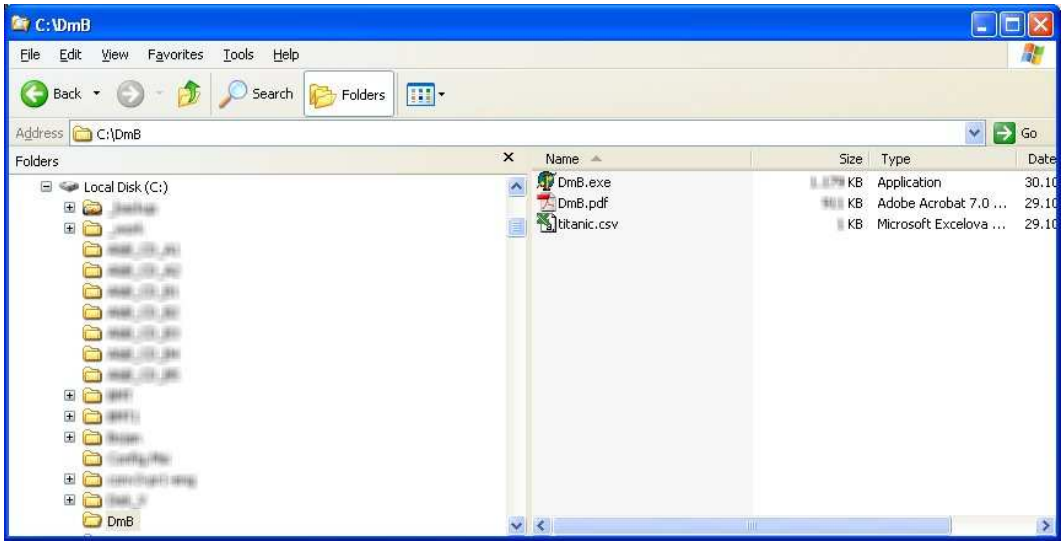
Instructions for use

May 2009

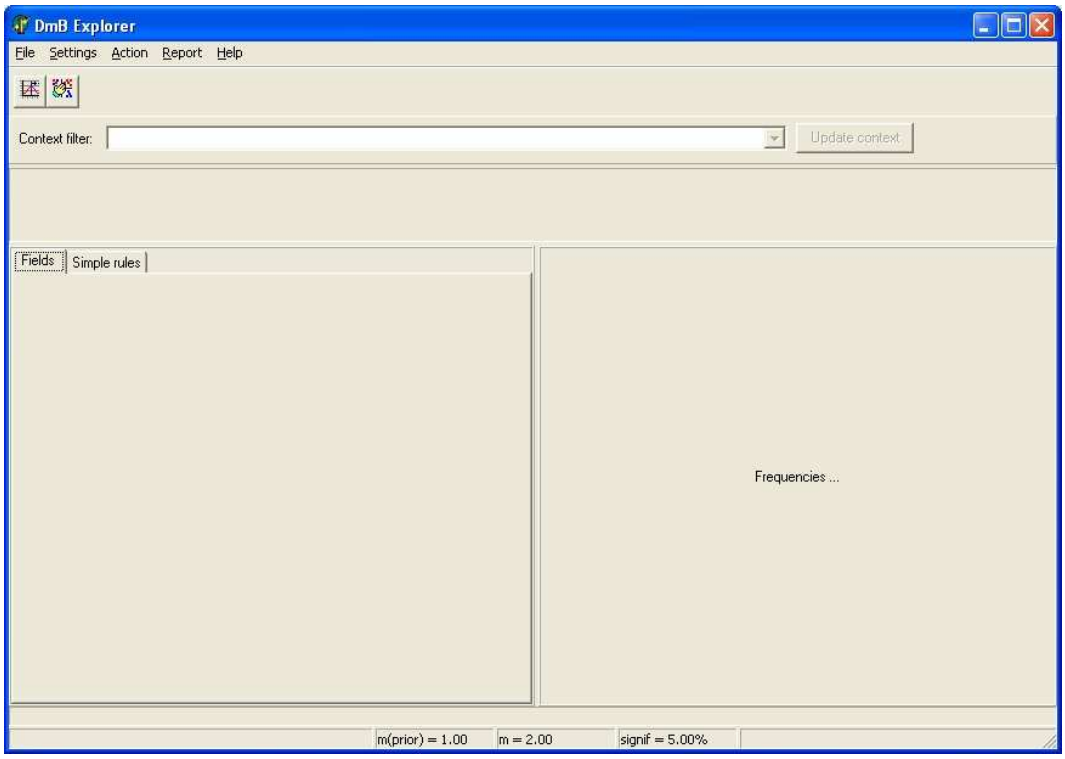
Typical sequence of steps for using the *DmB* program

1. Download the DmB.zip file and unzip its contents into a directory. For example, create directory \DmB on disk C: and copy the files the created directory.

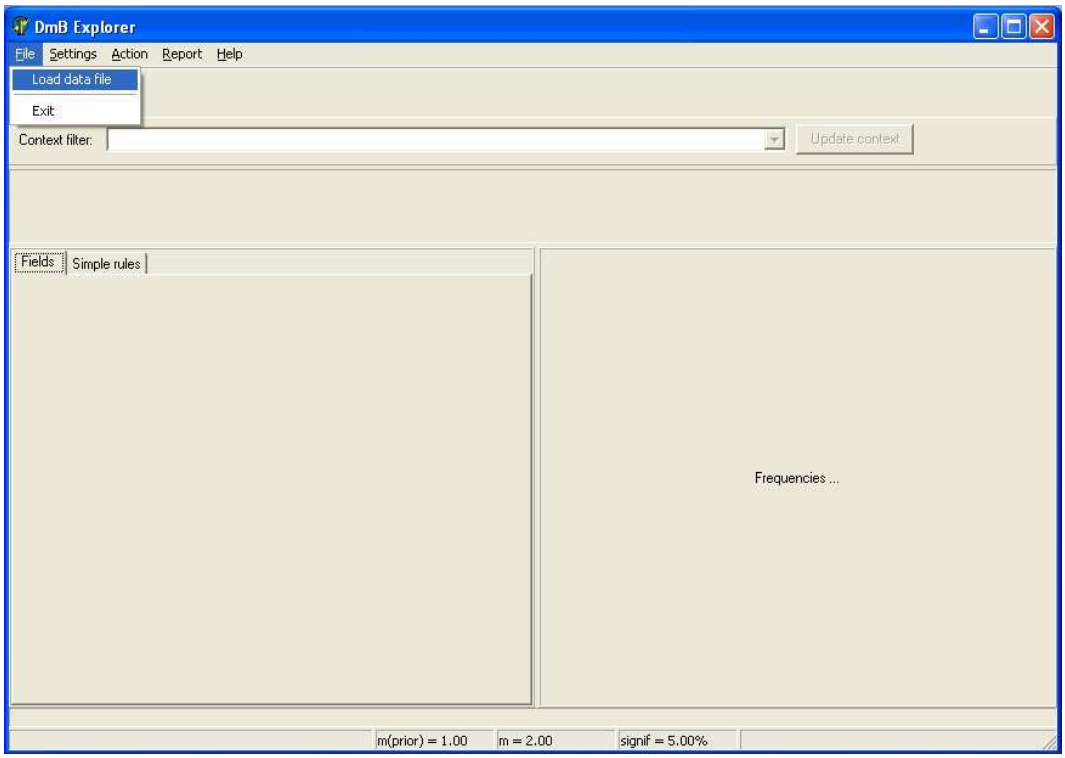
2. Open the created directory in Windows explorer. You should view three files: *DmB.exe* (executable file), *DmB.pdf* (this file) and *titanic.csv* (sample data file). Start the program by doubleclicking on DmB.exe icon.



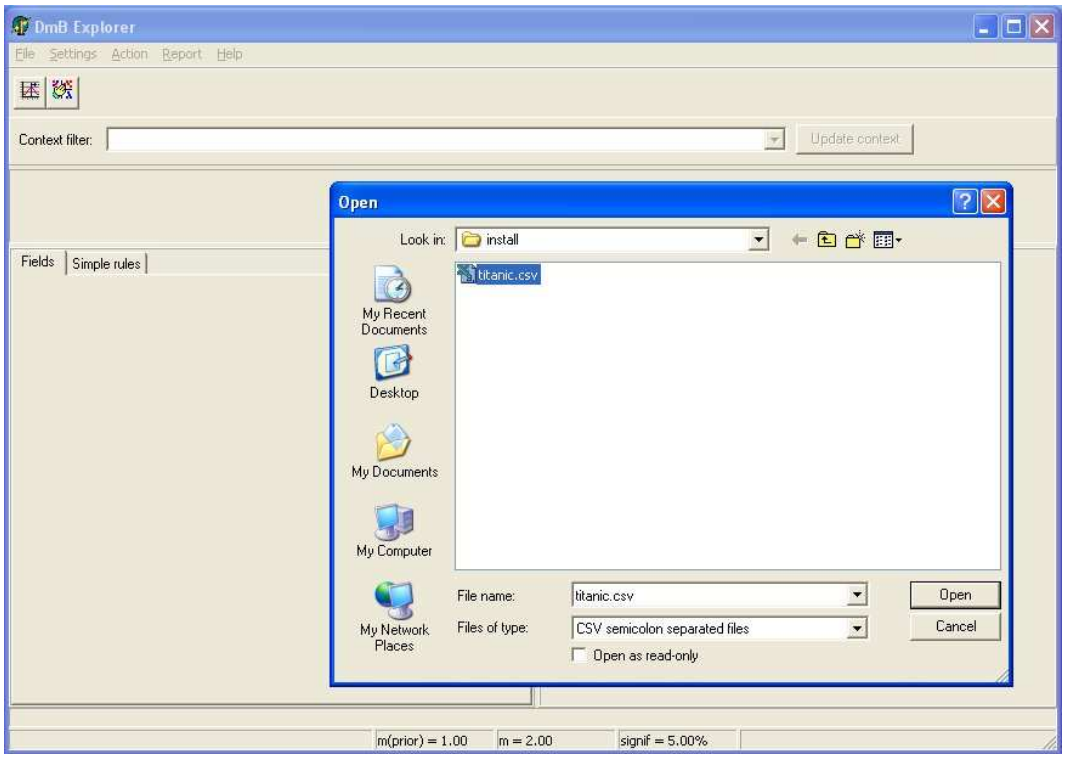
3. When you activate the DmB program, the following form should appear on your screen:



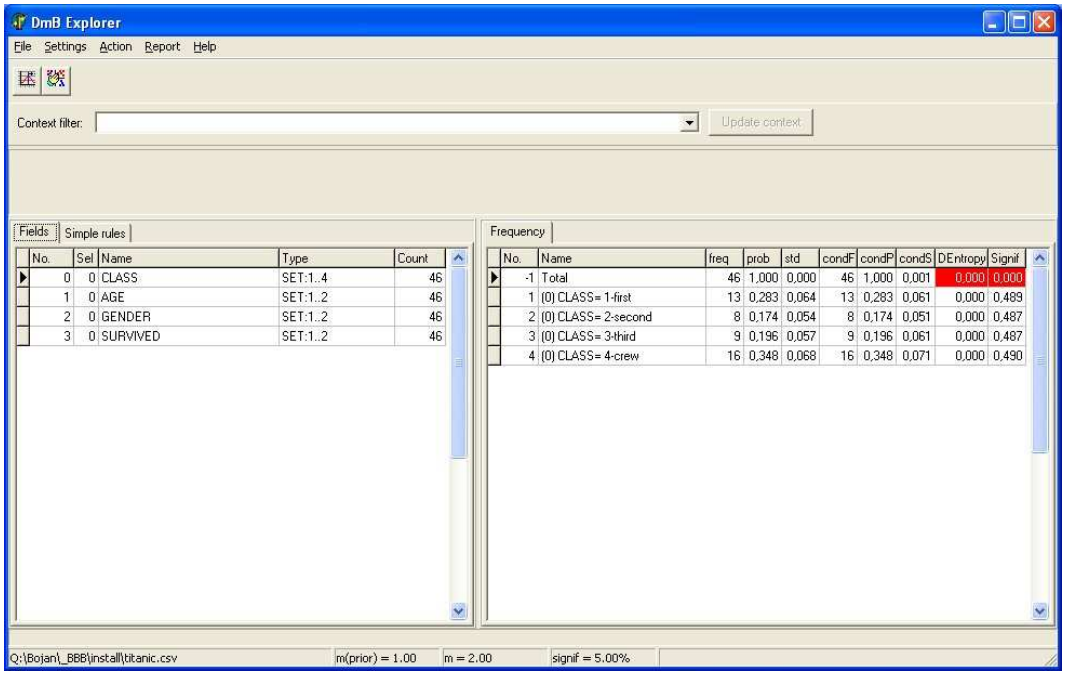
4. The first action is to **Load data file** from the **File** menu.



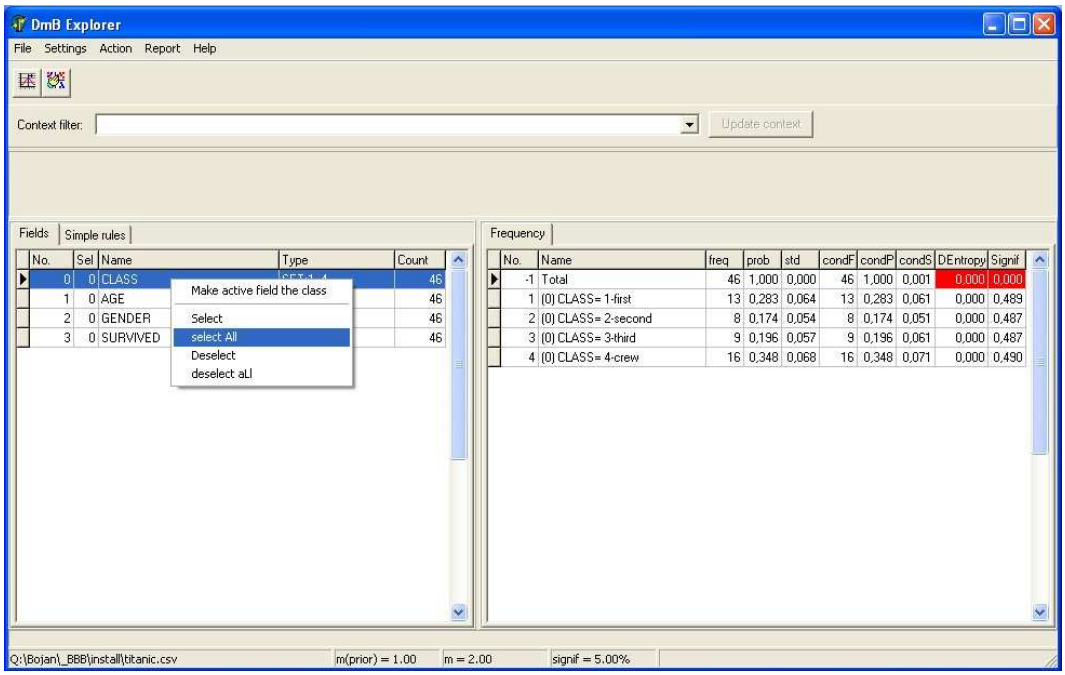
5. From the Open dialog select the data file you would like to use for the analysis. For demo purposes select *titanic.csv* and click **Open**.



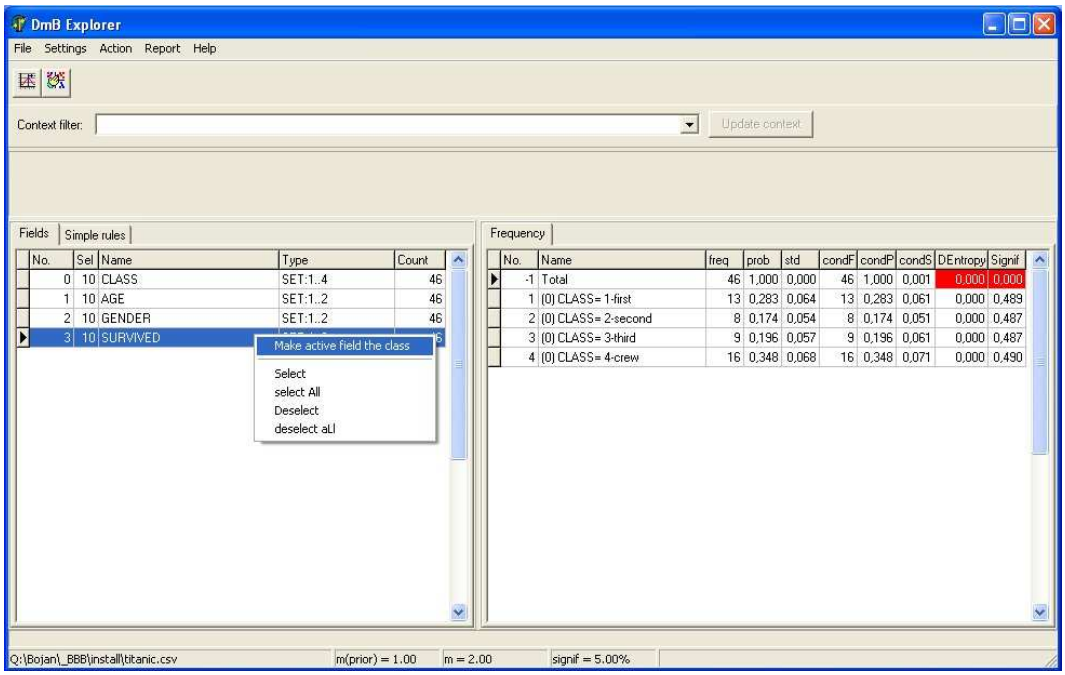
6. After you loaded the data file, the form should look like the one below. We can see that in the data file there are four fields (labeled 0 to 3). Note that the **Count** column in the Fields tab gives the total number of instances 46.



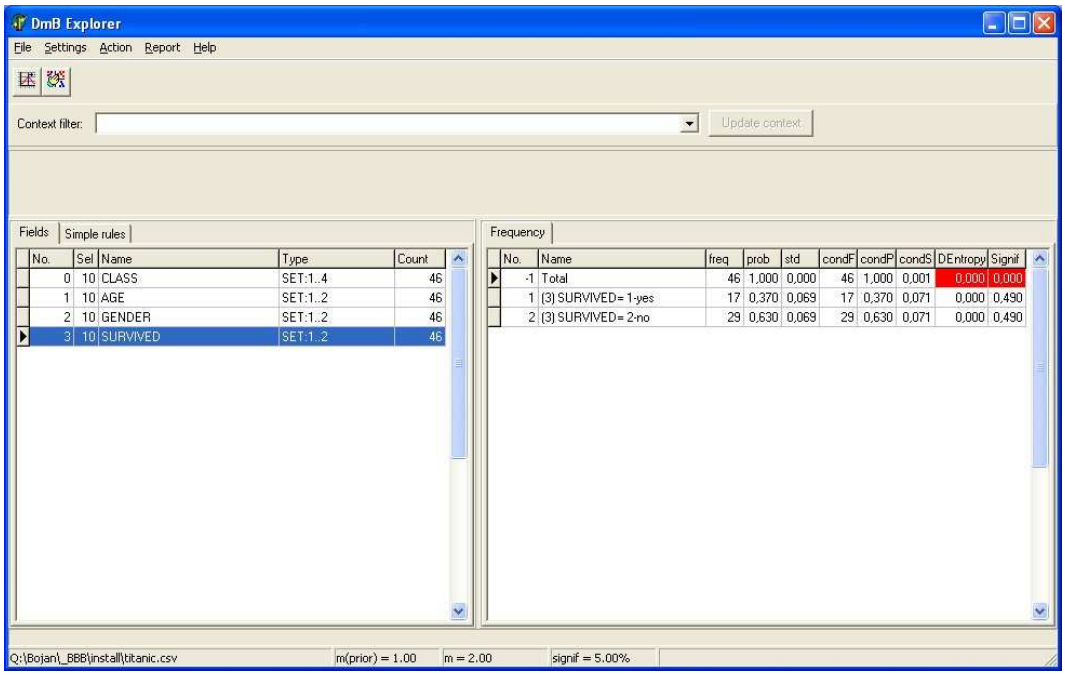
7. On the grid on the left hand side, click with the right mouse button to activate popup menu. Pick the **select All** fields, so that all the fields are selected for the analysis. **Fields** are sometimes referred to as *attributes*. The value 0 in the column **Sel** denotes that the field is **not** selected for the analysis, while the value 10 denotes the opposite – the field is selected. You can use any of the bottom four options to select or deselect the fields.



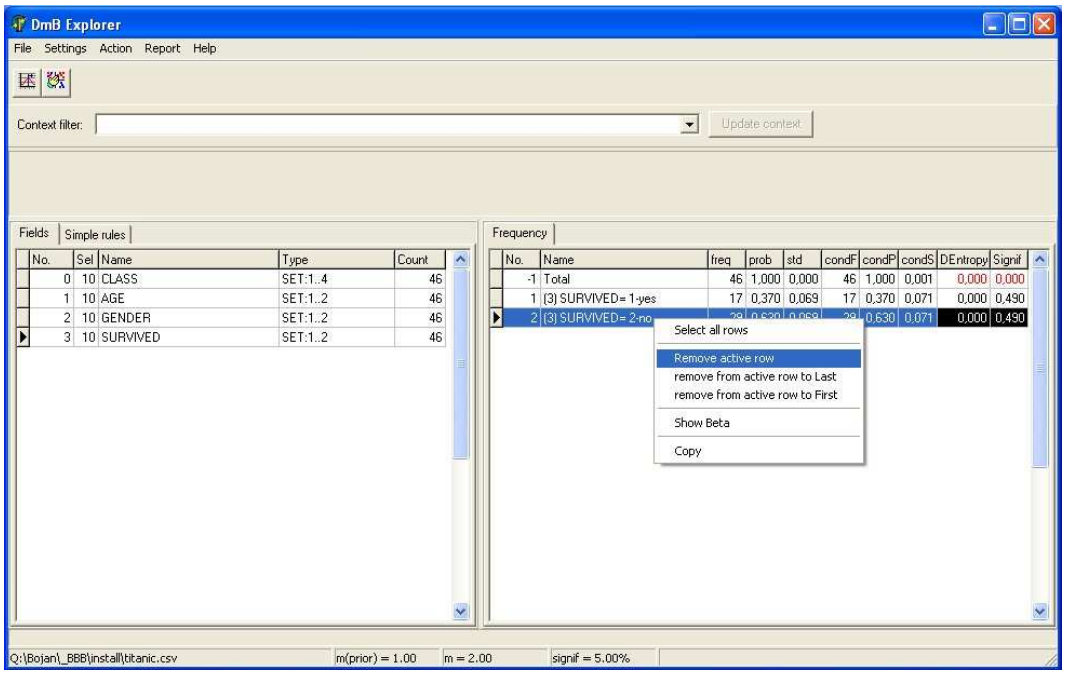
8. When you selected the fields for analysis, you have to denote which field will serve as a dependent field. We refer to the dependent field as **class**. In the analysis, conditional probabilities of a class values given the values of the selected fields are computed. In the data file *titanic.csv*, click on the last field (SURVIVED) to select it (selected fields are shown with blue background), then activate pop-up menu with the right-click on the selected field and select menu option **Make active field the class**.



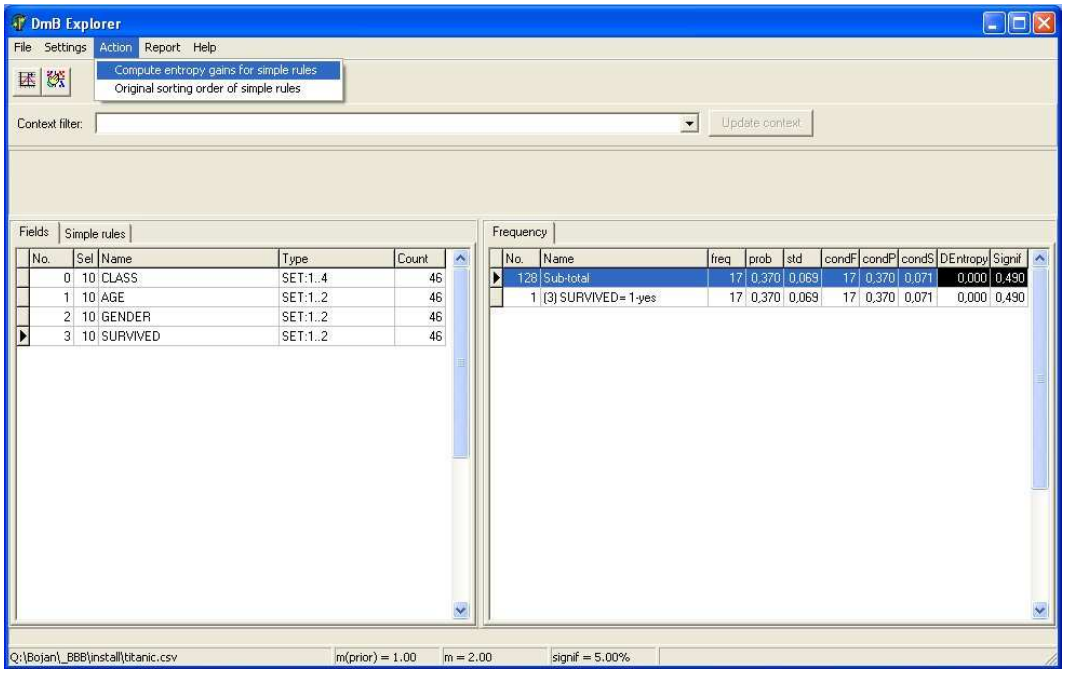
9. Observe the change in the right tab Frequency. Now, the class values appear as rows in the matrix. At this point let us describe the first five columns of the Frequency matrix. The first column is labeled **No.** and contains index of a class value. Note that the value –1 stands for **Total** row. **Name** contains description of a class value. Columns **freq**, **prob** and **std** stand for frequency, probability and standard deviation respectively.



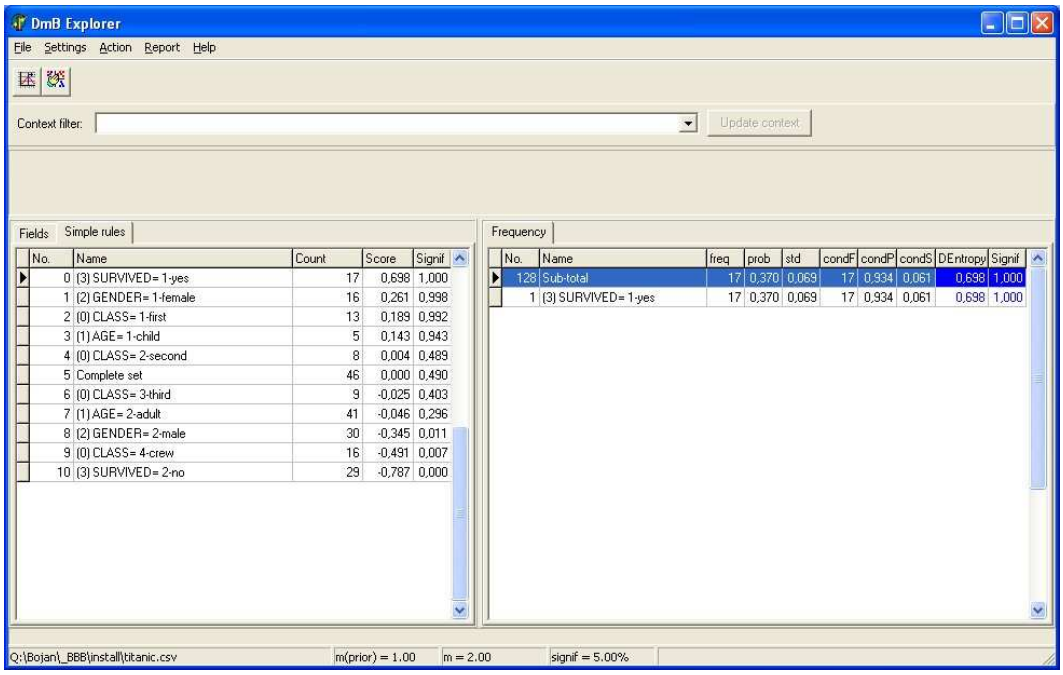
10. Suppose that in analyzing the *titanic.csv* data file we are only interested in probabilities of surviving the Titanic sinking. So, we have to hide the value **SURVIVED=2-no** from the **Frequency** tab. We accomplish this by right-clicking on the second class value and selecting the option **Remove active row** from the pop-up menu.



11. From the main menu we select **Action** and **Compute entropy gains for simple rules**. This procedure computes conditional probabilities for simple rules (the tab in the background of the bottom-left side of the form).



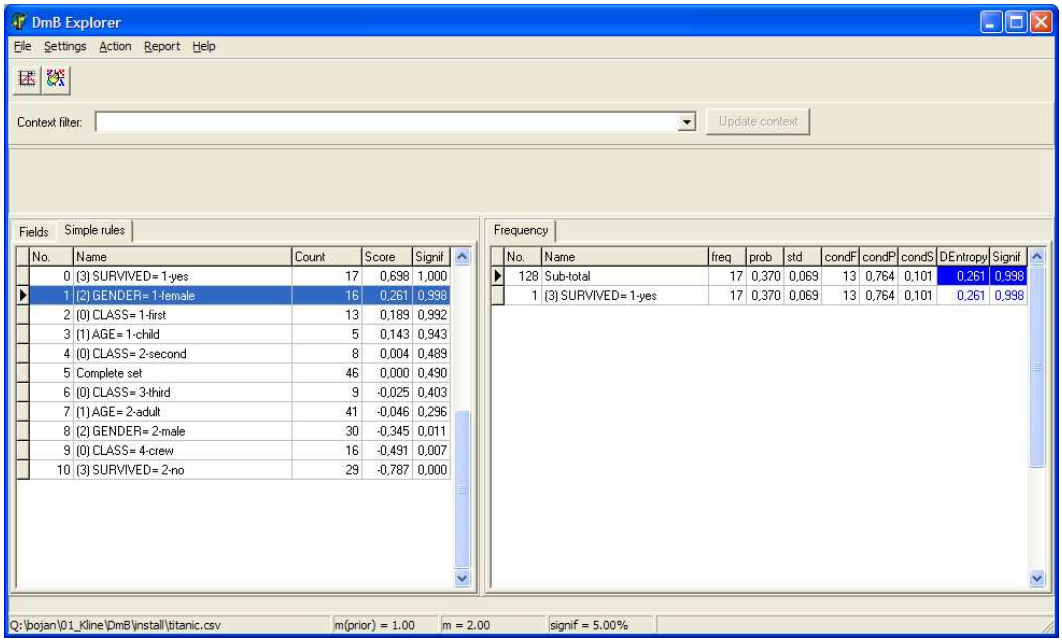
12. Then we select the tab **Simple rules**. Simple rules are made of field-value pairs. For each field-value pair we see **Count** denoting the number of occurrences of this field-value pair in the data file. The next two columns are **Score** and **Signif**, which we will explain in the next step.



No.	Name	Count	Score	Signif
0	(3) SURVIVED= 1-yes	17	0,698	1,000
1	(2) GENDER= 1-female	16	0,261	0,998
2	(0) CLASS= 1-first	13	0,189	0,992
3	(1) AGE= 1-child	5	0,143	0,943
4	(0) CLASS= 2-second	8	0,004	0,489
5	Complete set	46	0,000	0,490
6	(0) CLASS= 3-third	9	-0,025	0,403
7	(1) AGE= 2-adult	41	-0,046	0,296
8	(2) GENDER= 2-male	30	-0,345	0,011
9	(0) CLASS= 4-crew	16	-0,491	0,007
10	(3) SURVIVED= 2-no	29	-0,787	0,000

No.	Name	freq	prob	std	condF	condP	condS	DEntropy	Signif
128	Sub-total	17	0,370	0,069	17	0,934	0,061	0,698	1,000
1	(3) SURVIVED= 1-yes	17	0,370	0,069	17	0,934	0,061	0,698	1,000

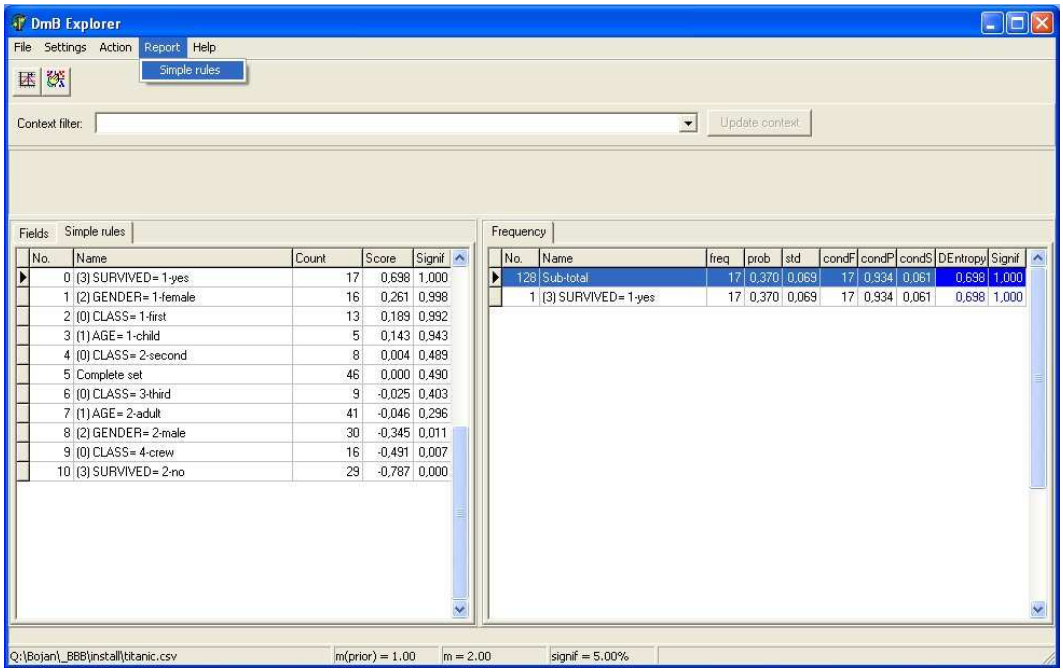
13. Next we click on the simple rule GENDER=1-female. **Count** column states that there are 16 females in the whole data set. Now it is the right time to explain the last five columns in the **Frequency** tab. But first, let us recap the **freq**, **prob** and **std** columns. We can see that in total there are 17 survivors out of 46 instances, resulting in estimated *m*-probability of 0,370 and standard deviation of 0,069. For the theoretical background of *m*-probability estimate look in the references section. The last five columns in the **Frequency** tab are **condF**, **condP**, **condS**, **DEntropy** and **Signif**. The column **condF** stands for conditional frequency, indicating that there are 13 survivors that are also females. **condP** is conditional probability of surviving given that you are female, calculated according to the *m*-estimate. The actual figure is 0,764. Observe that females have much greater chance of surviving than passengers on average. **condS** gives standard deviation of the *m*-estimate of probability. **DEntropy** gives a measure of information (*Difference of entropy*) about the chance of survival that is received by the fact that a passenger is female. It is measured in bits; positive value means information in favor of the class, negative value means information against the class. Difference of entropy can be referred to as *Information gain*. **Signif** is significance measure that gives the probability that the conditional probability is greater than unconditional probability. In our case we have standard deviation of 0,101. The amount of information received in favor of surviving by the fact that a passenger is female in 0,261 bits. The probability that conditional probability of 0,764 is greater than unconditional probability of 0,370 is 0,998. Note that this is taken as a measure of significance of the result.



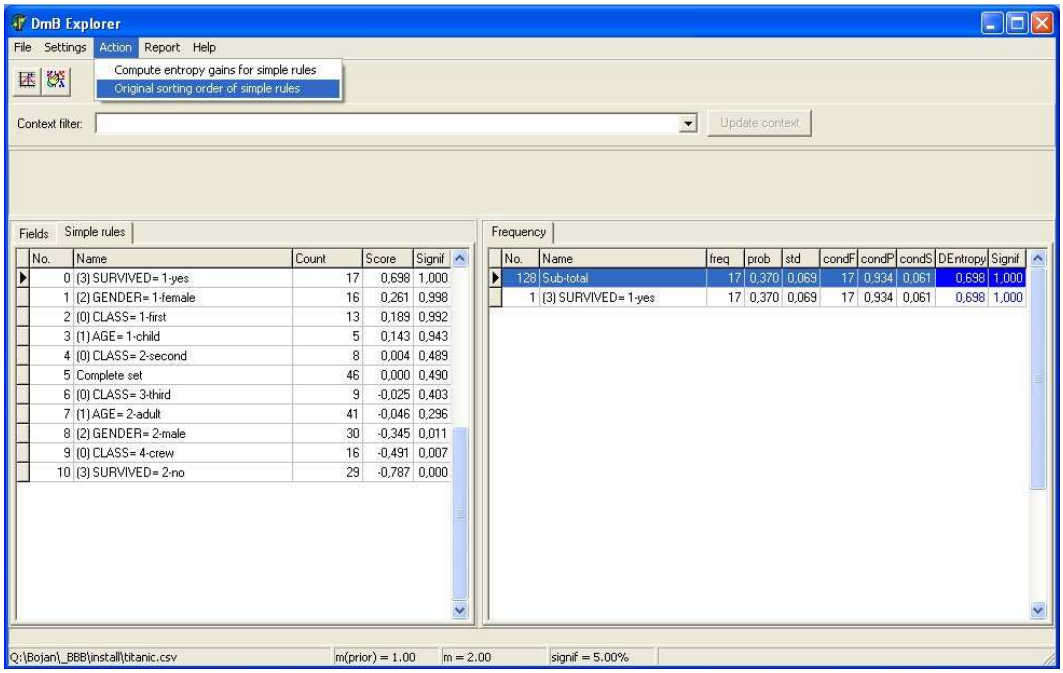
No.	Name	Count	Score	Signif
0	(3) SURVIVED= 1-yes	17	0,698	1,000
1	(2) GENDER= 1-female	16	0,261	0,998
2	(0) CLASS= 1-first	13	0,189	0,992
3	(1) AGE= 1-child	5	0,143	0,943
4	(0) CLASS= 2-second	8	0,004	0,489
5	Complete set	46	0,000	0,490
6	(0) CLASS= 3-third	9	-0,025	0,403
7	(1) AGE= 2-adult	41	-0,046	0,296
8	(2) GENDER= 2-male	30	-0,345	0,011
9	(0) CLASS= 4-crew	16	-0,491	0,007
10	(3) SURVIVED= 2-no	29	-0,787	0,000

No.	Name	freq	prob	std	condF	condP	condS	DEntropy	Signif
128	Sub-total	17	0,370	0,069	13	0,764	0,101	0,261	0,998
1	(3) SURVIVED= 1-yes	17	0,370	0,069	13	0,764	0,101	0,261	0,998

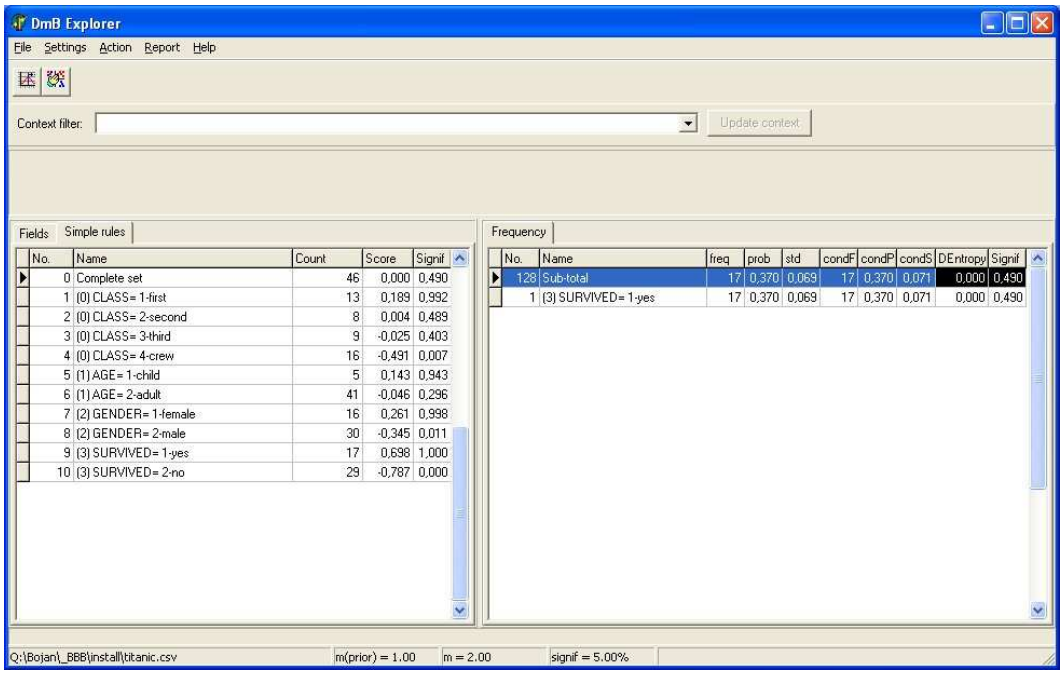
14. Select Report and Simple rules from the main menu to generate report to the DmBGlobalRep.txt file.



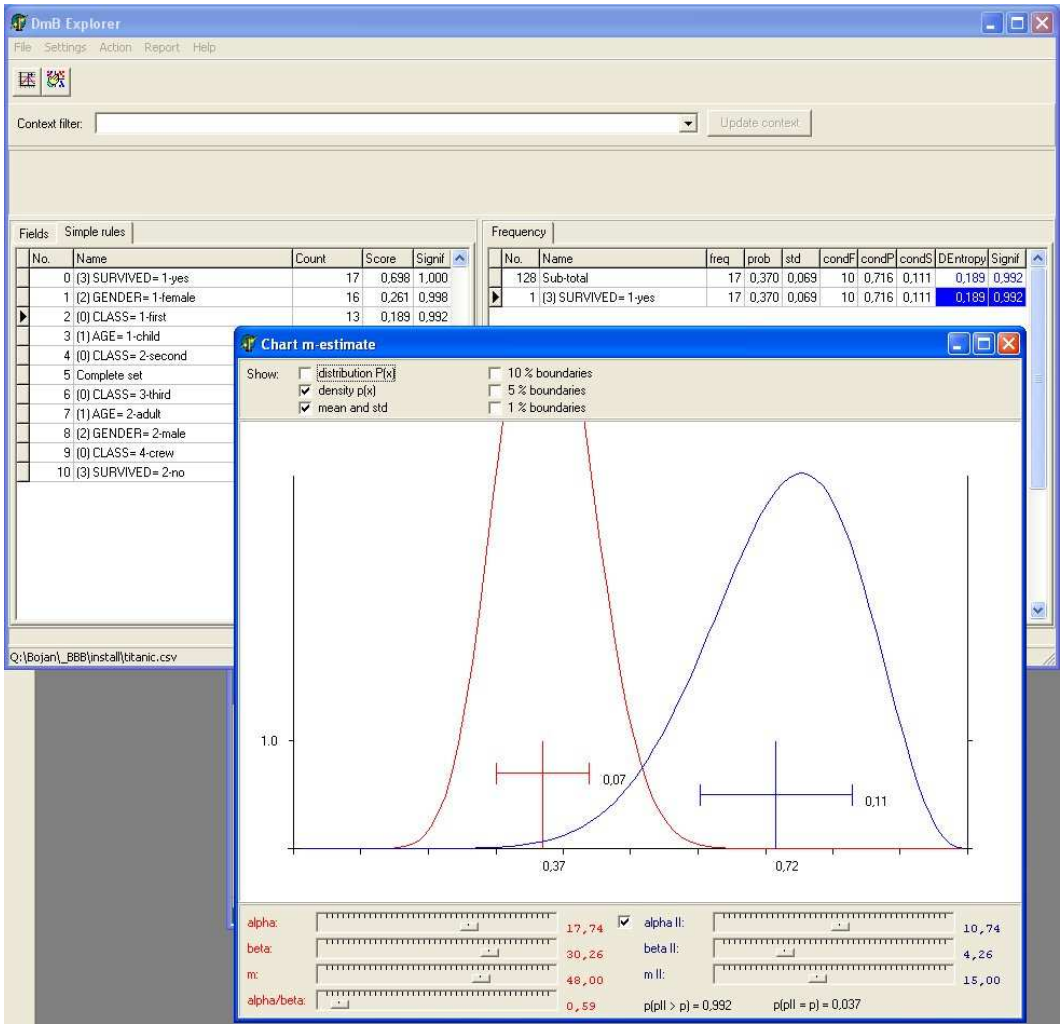
15. Note that the simple rules are sorted according to the significance level of rules (**Signif**). By selecting **Action** and **Original sorting order of simple rules** the simple rules are sorted as they were generated at the beginning of the execution of the program.



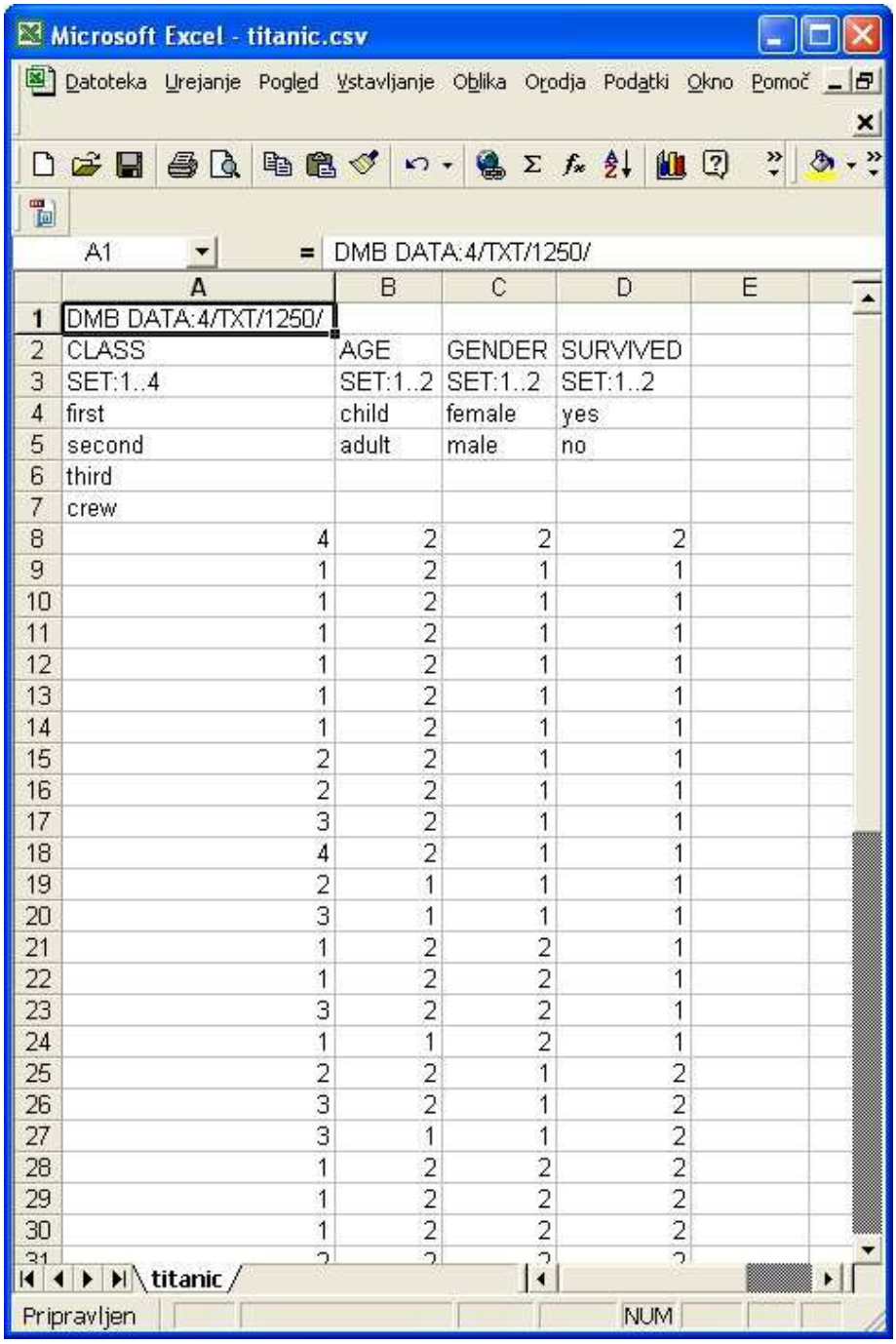
16. The result of the original sorting order is shown in the picture below.



17. To visualize the underlying probability distributions (Beta) right-click on the class value and select **Show Beta** option form the pop-up menu. Select **density p(x)** and **mean and std** for the similar effect as below. Note that the selected simple rule for the example below is CLASS=1-first.



Ad. 1. The data file has the form of a *.csv* file (comma-separated values). It can be edited in Excell or equivalent utility. The structure is given in the picture below. Grammatical rules are given in Ad. 2.



	A	B	C	D	E
1	DMB DATA:4/TXT/1250/				
2	CLASS	AGE	GENDER	SURVIVED	
3	SET:1..4	SET:1..2	SET:1..2	SET:1..2	
4	first	child	female	yes	
5	second	adult	male	no	
6	third				
7	crew				
8	4	2	2	2	
9	1	2	1	1	
10	1	2	1	1	
11	1	2	1	1	
12	1	2	1	1	
13	1	2	1	1	
14	1	2	1	1	
15	2	2	1	1	
16	2	2	1	1	
17	3	2	1	1	
18	4	2	1	1	
19	2	1	1	1	
20	3	1	1	1	
21	1	2	2	1	
22	1	2	2	1	
23	3	2	2	1	
24	1	1	2	1	
25	2	2	1	2	
26	3	2	1	2	
27	3	1	1	2	
28	1	2	2	2	
29	1	2	2	2	
30	1	2	2	2	
31	2	2	2	2	

Ad. 2.

<data file>:- <header line><nl><filed names line><nl><field types line><nl><field values><nl><actual values for instances>

<nl>:- new line

<header line>:- <dmb data stamp><colon><maximal number of field values><slash><content indicator><slash><code page>

<colon>:- ":"
<slash>:- "/"
<dmb data stamp>:- "DMB DATA"
<maximal number of field values>:- max number of field values
<content indicator>:- "TXT"
<code page>:- "1250"

<field names line>:- <empty> | <field name><field separator><field names line>

<empty>:- empty
<filed name>:- name of a field
<field separator>:- ";"

<field types line>:= <empty> | <field type><field separator><field types line>

<field type>:- <set type> | <int type> | <real type>

<set type>:- <set><colon><min value><two dots><max value>

<set>:- "SET"
<two dots>:- ".."
<min value>:- minimal value
<max value>:- maximal value

<field values>:- names (labels) of filed values

<actual values for instances>:- actual filed values for intances