CAMPUS
DE EXCELENCIA
INTERNACIONAL

Master Universitario en Ciencia de Datos

Practical Application 2

Probabilistic Graphical Models

El Abbassi Widad

January 7th,2019

I. Introduction

This article is revolving around an application that was developed in Yugoslavia for admission in public school systems to evaluates, classifies and ranks options to accept students and then bring a logical answer to the parents.

The school authorities established a pilot project at a particular school with the aim to develop a methodology for selecting applicants, which required the establishment of an expert task force team consisting of 13 members including two decision theory experts, one of them selected as chairman of the group and both given the technical responsibility of the project.

The main tool used in developing the decision support system to rank nursery school applications was DECMAK (Bohanec & Rajkovi˘c 1987). This is an expert system shell specialized for multi- attribute decision problems, i.e. problems where several alternatives (also called options) are to be ranked according to their quality. The quality depends on a particular set of attributes.

II. Problem description

The problem in this article is the selection of applicants for public nursery schools, so that we could provide an explanation along with the rejected applications. The selection is supported by an expert system which evaluates, classifies and ranks applications. According to the article " An application for admission in public school systems " published by [Vladislav Rajkovic, Marko Bohanec and Manuel Olave] The Decision rules are entered into the knowledge base in the If-Then form, for example:

IF Social conditions of the family are problematic,

AND Health conditions of the family allow acceptance of a child,

THEN a priority acceptance of the child is recommended.

In this article, the goal is to learn a network from a dataset (structure and parameters), and then preform some inference.

III. Methodology

1. Analysis of the Data

NURSERY dataset is a real-world model developed to rank applications for nursery schools (Olave et al. 1989), this dataset contains 8 features that on the basis of them we have to make a decision if the child will be admitted to the public school system. The choice of this particular dataset is due to its popularity among the machine learning researchers and also in order to avoid information loss with discretization since it doesn't contain any continuous variables.

There are three main categories of attributes:

1. Occupation of parents and characteristics of the nursery.

2. Family structure and financial standing,

3. Social and health status of the family.

These categories are represented by internal nodes as follows:

NURSERY Evaluation of applications for nursery schools
. EMPLOY Employment of parents and child's nursery
. . parents Parents' occupation
. . has_nurs Child's nursery
. STRUCT_FINAN Family structure and financial standings
. . STRUCTURE Family structure
. . . form Form of the family
. . . children Number of children
. . housing Housing conditions
. . finance Financial standing of the family
. SOC_HEALTH Social and health picture of the family
. . social Social conditions
. . health Health conditions

2. Attributes Values

Final evaluation value of the application**:** (1) acceptance is not recommended, (2) acceptance is recommended, (3) acceptance is very much recommended, (4) priority acceptance is recommended and (5) special priority acceptance is recommended.

Occupation of parents and child's nursery**:** (1) convenient, (2) less convenient, (3) inconvenient and (4) critical.

- Parents' occupation: (1) usual, (2) pretentious, (3) of great pretension.
- Child's nursery: (1) proper, (2) less proper, (3) improper, (4) critical, (5) very critical.

Family structure and financial standing: (1) convenient, (2) inconvenient and (3) critical.

Family structure**:** (1) less critical, (2) critical and (3) very critical.

- Form of the family: (1) complete family, (2) completed family, (3) incomplete family, (4) foster family.
- Number of children: number of preschool and school-age children.
- Housing conditions: (1) convenient, (2) less convenient and (3) critical.
- Financial standing of the family: (1) convenient, (2) inconvenient.

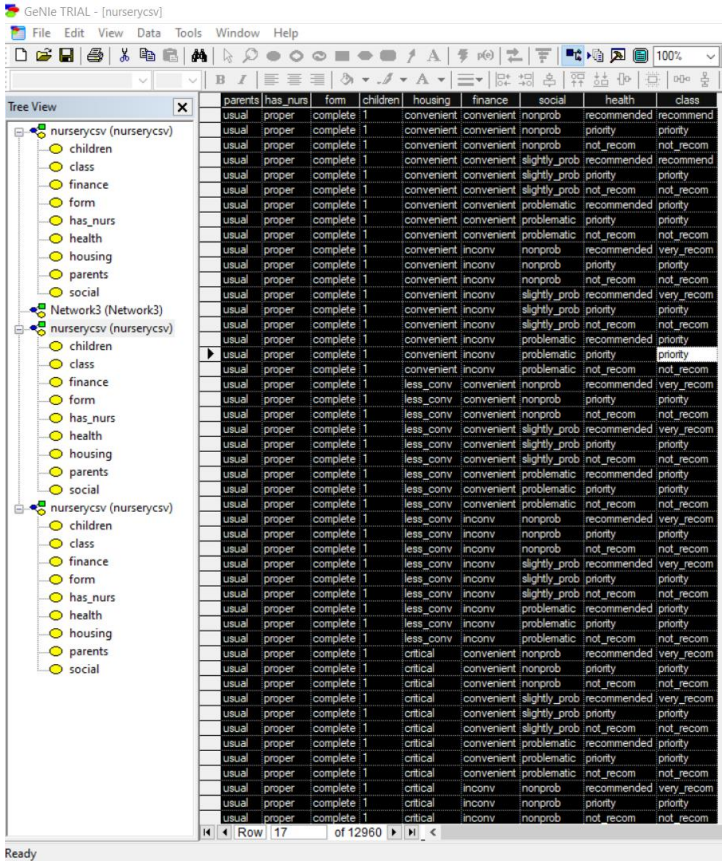
Social and health picture of the family: (1) acceptance is not recommended, (2) acceptance is recommended and (3) priority acceptance is recommended.

- Social conditions: (1) non-problematic, (2) slightly problematic and (3) problematic.
- Health conditions: (1) acceptance is not recommended, (2) acceptance is recommended and (3) priority acceptance is recommended.

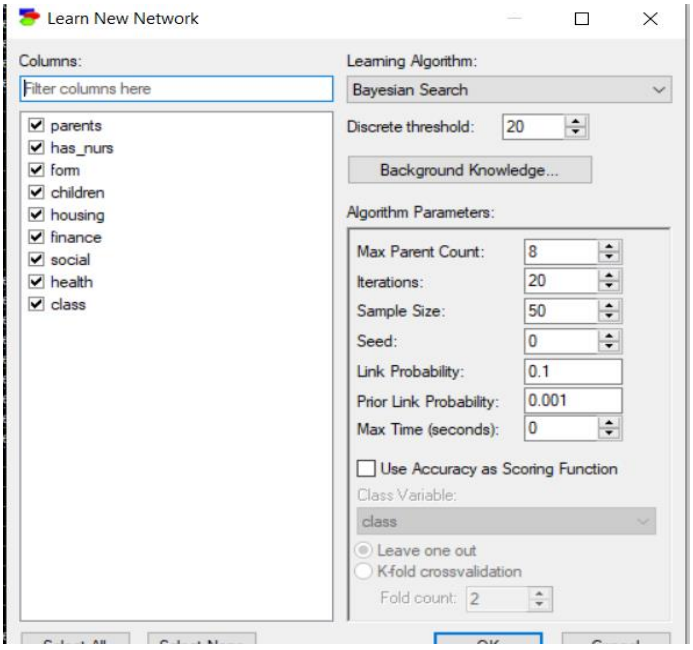
IV. Results

1. Learning Bayesian Network:

As a first step in this work, we will begin by loading the Nursery Dataset in the BayesF Fusion software after making sure that no continuous variables are left without discretizing no and missing values exist:

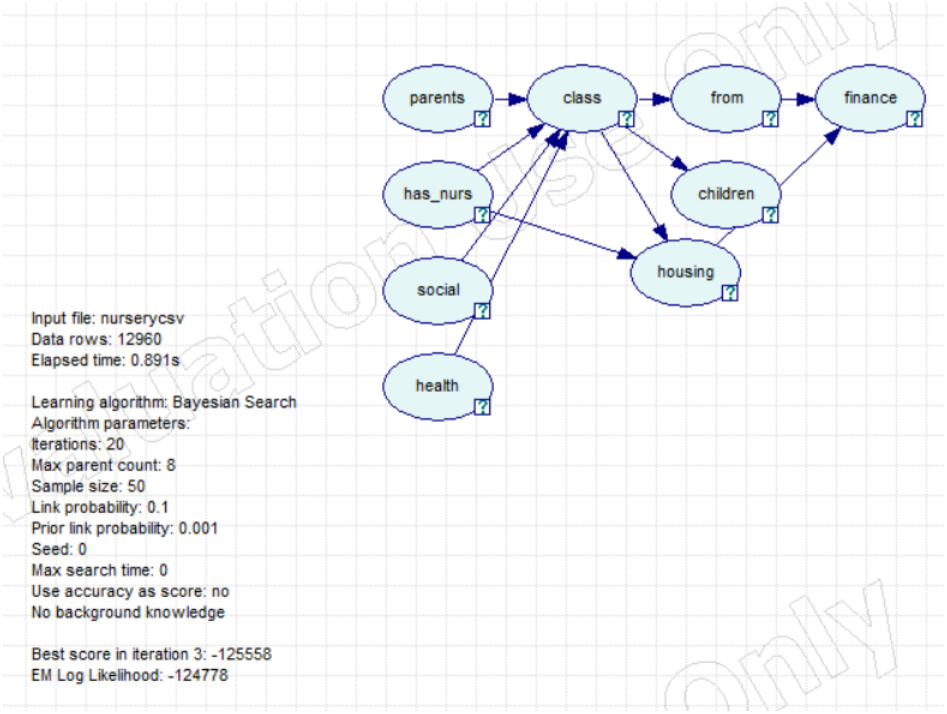


After loading the data, we will set parameters for learning as shown below :

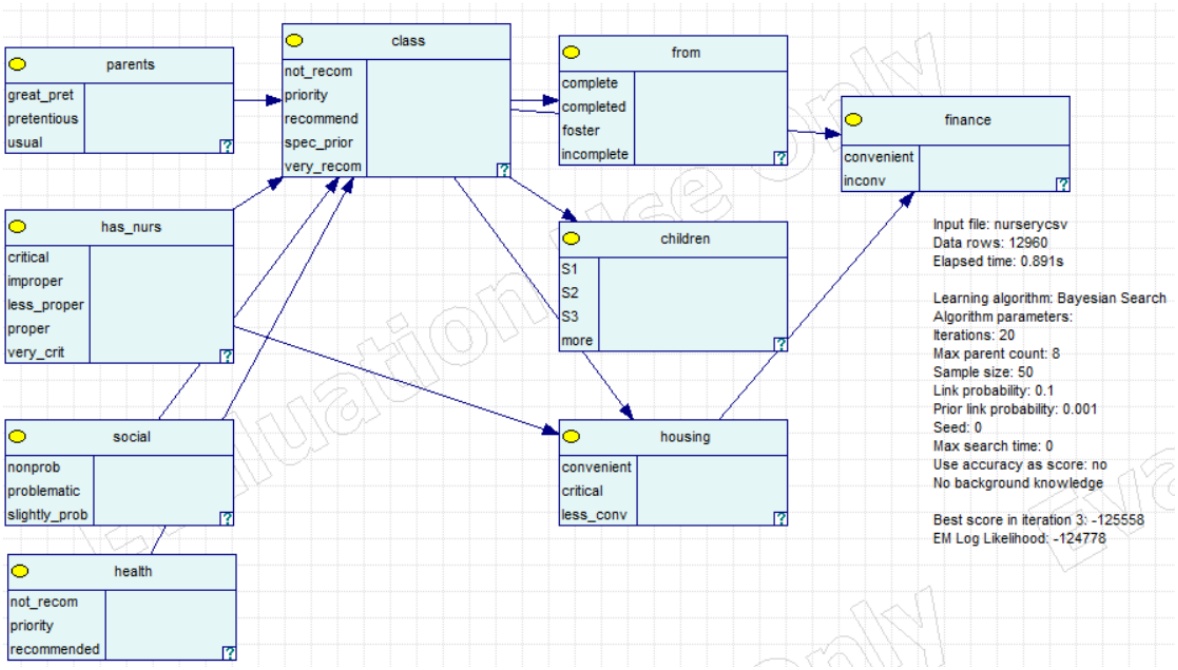


NOTE : we can also apply forcing or forbidding edges in in background knowledge.

As a result , we got our network structure.



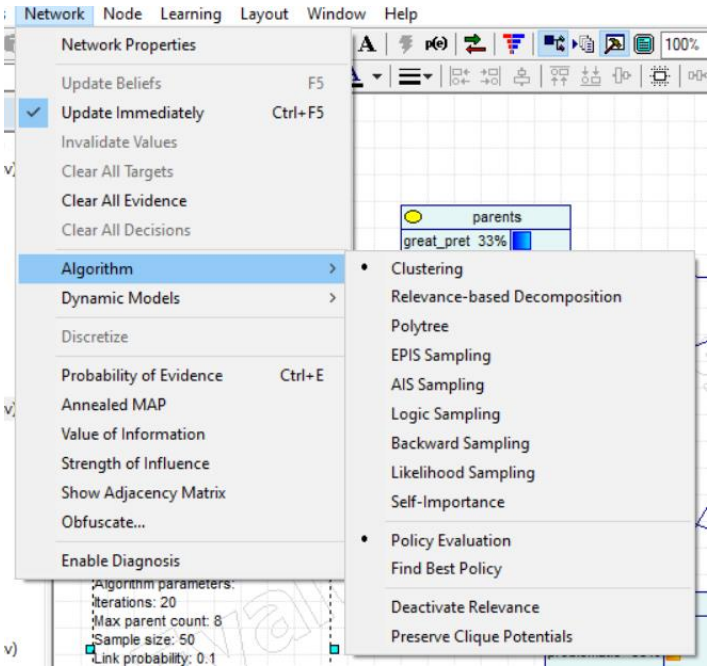
To make it more understandable, we will reorganize it and change the layout to have a better observation.



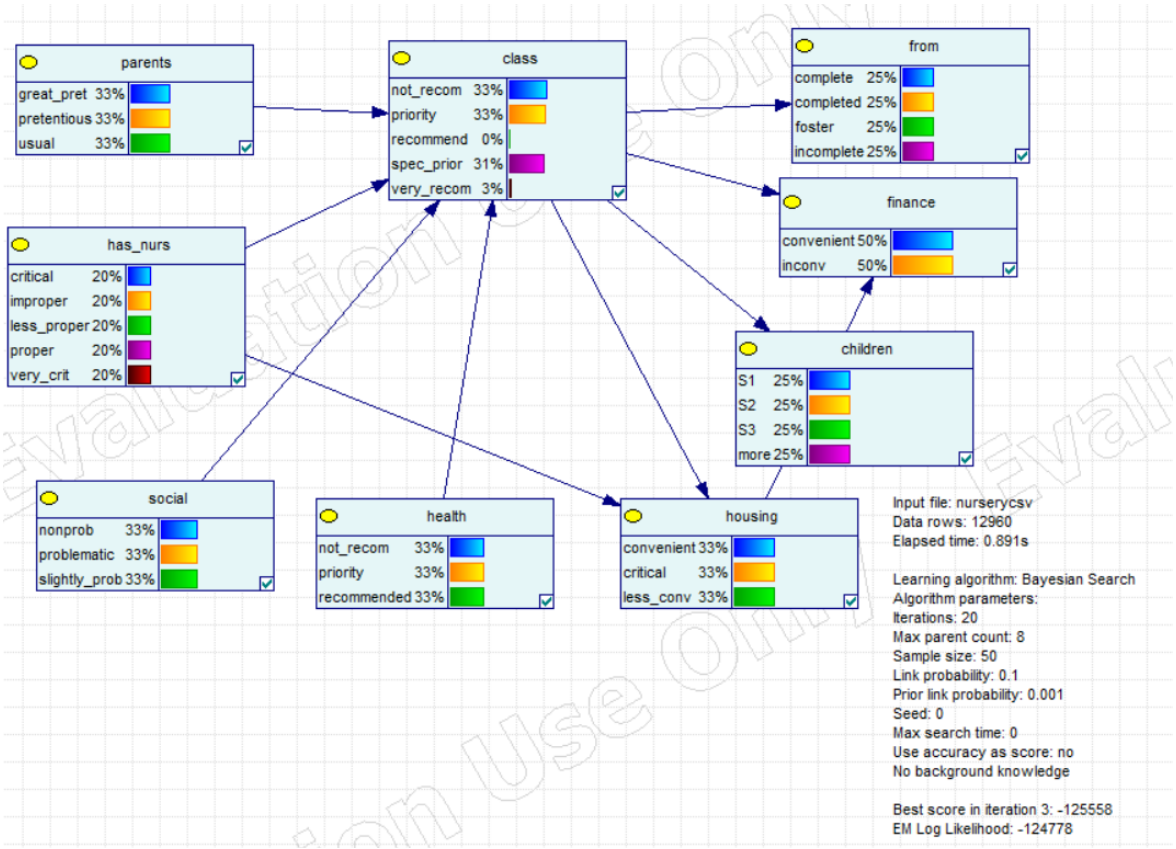
The learning algorithm chosen here is the Bayesian Search with no background knowledge.

2. Exact Inference:

For this part, we choose Clustering as network Algorithm.



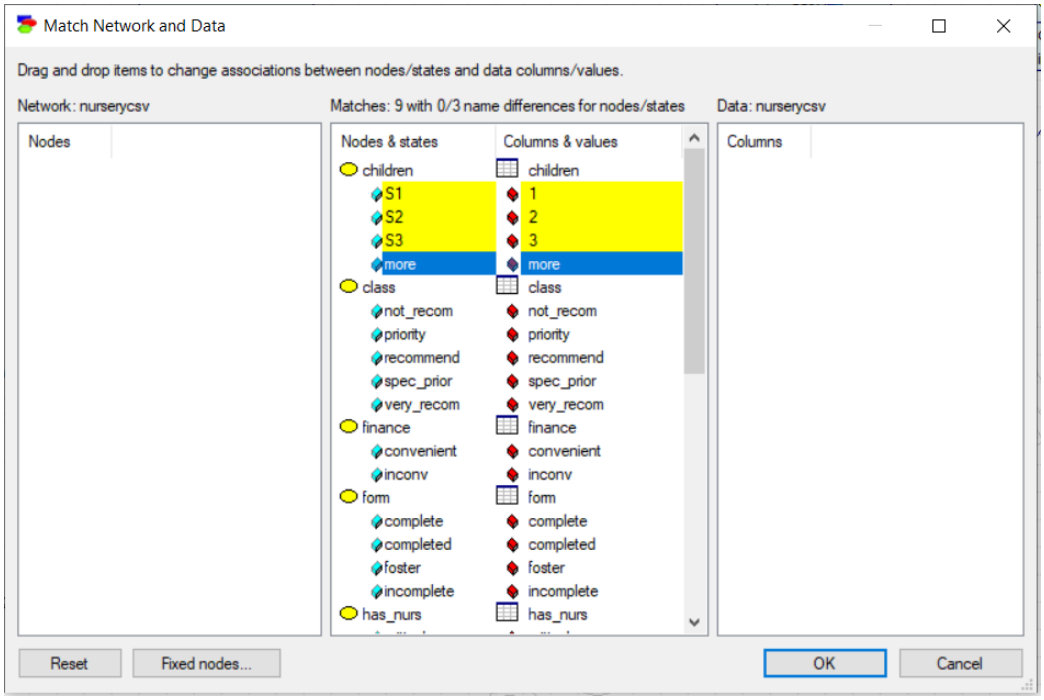
Then we use - immediately updates the model - so as soon as any change, observation, or control is made to the model, it will be updated.



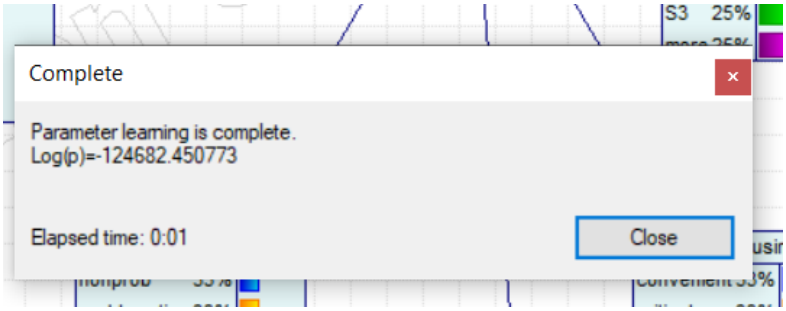
We have as a result a Bayesian network that learned from our dataset (we get a marginalized probabilities for all the nodes).

- ✓ **Learn the parameters:**

This will invoke the Match Network and Data dialog that serves to create a mapping between the variables defined in the network and the variables defined in the data set.



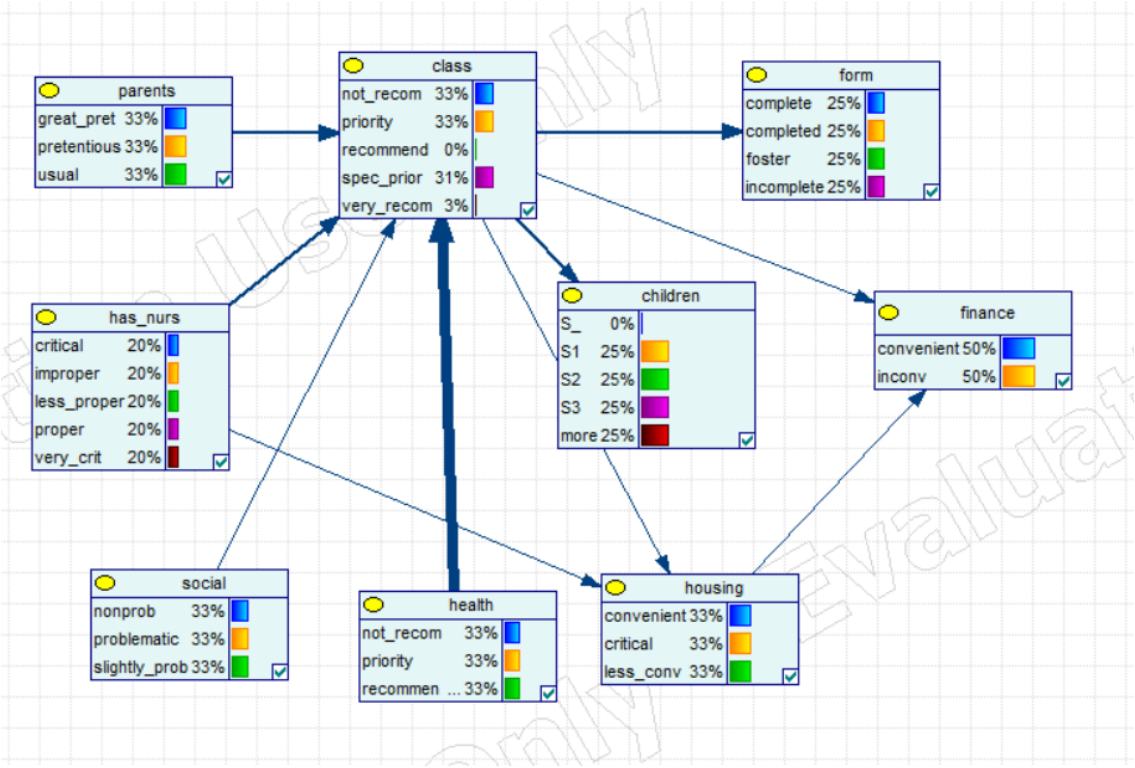
After that, the EM algorithm updates the network parameters following the options chosen and comes back with the following result:



This value measures the fit of the model to our Nursery data.

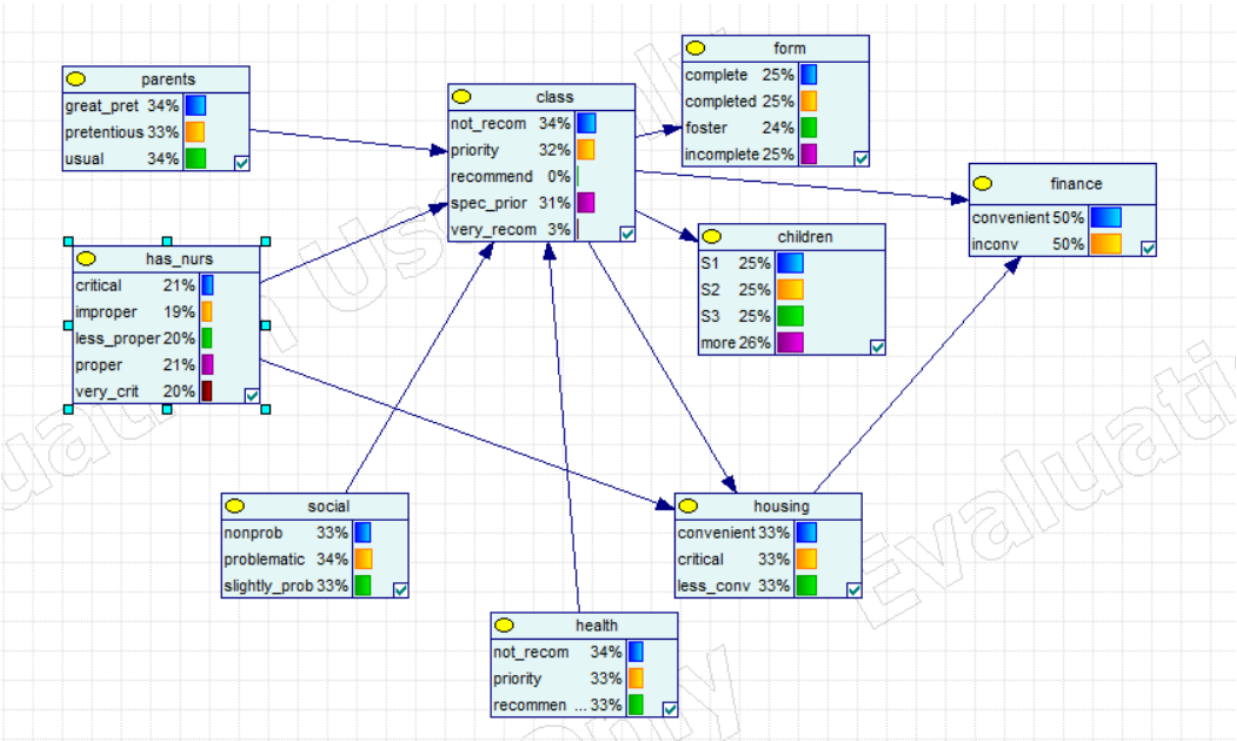
- ✓ **Strength of influence:**

In Genie software, we have the option to observe the strength of relationships between the variables.



3. Approximate Inference

Here we change the algorithm applied in the network tab to Logic Sampling and a slight change can be seen in the variables probabilities.

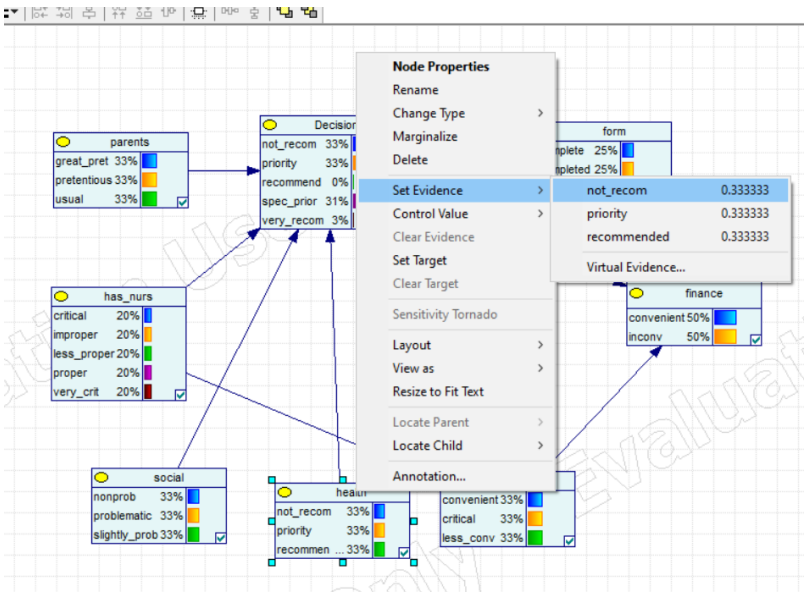


✓ **Inference (Manually):**

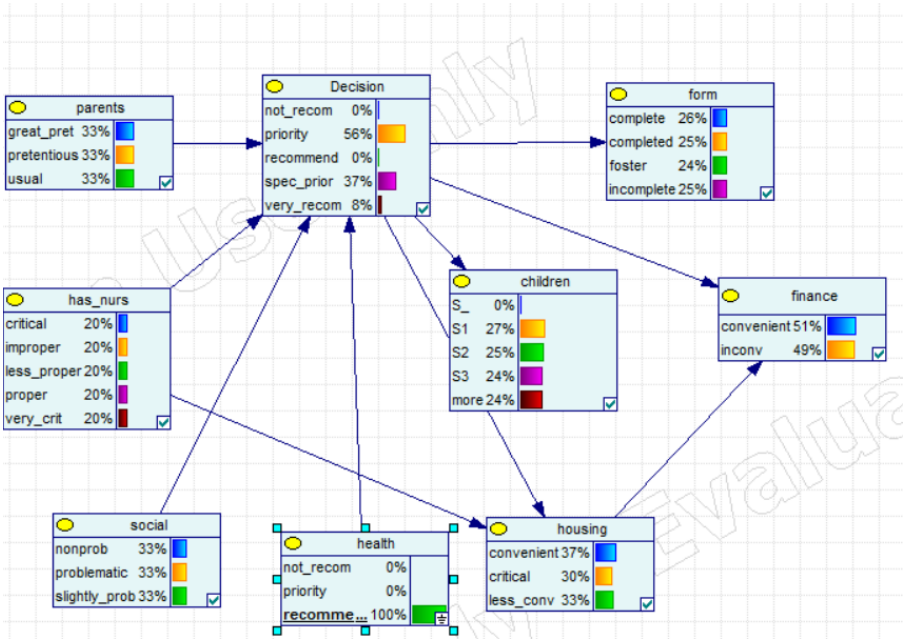
As a test, we will try to observe the input variables to see the impact of some of them on the decision regarding the application acceptance, and to do so we will follow these steps: Set evidence, Update belief and then check updated probability.

⇨ **Setting evidence:** health to “recommended”

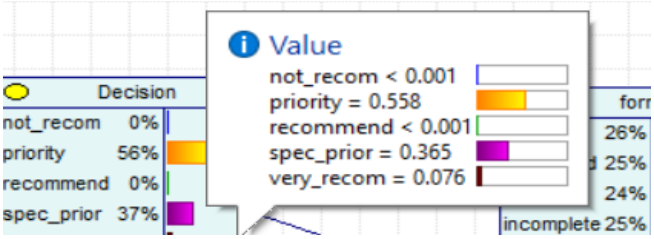
In the node menu, we choose Set Evidence submenu:



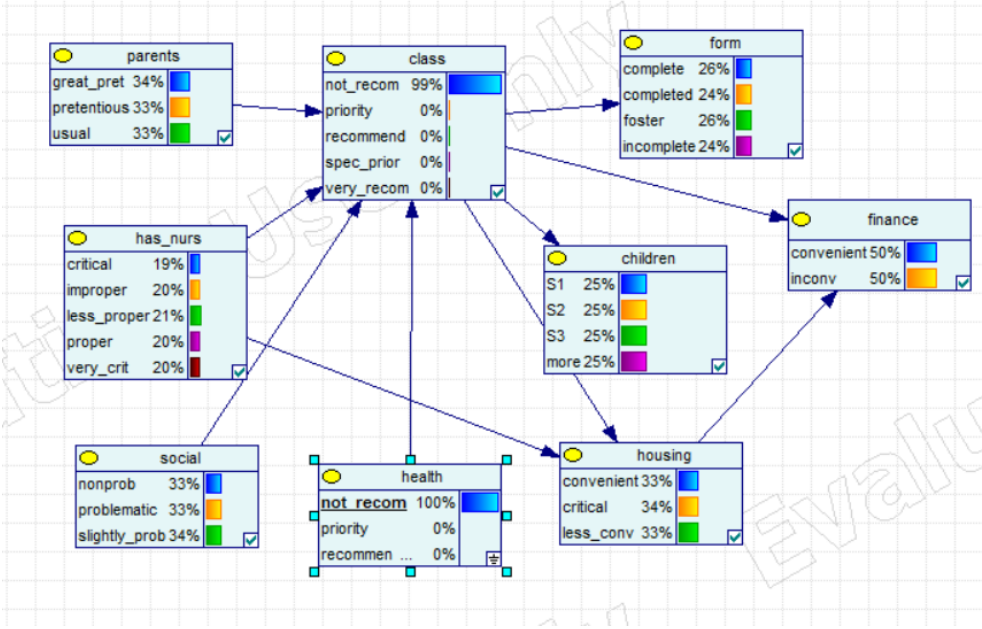
This updates the probability distributions in light of the observed evidence, and as a result we can see the impact of this:



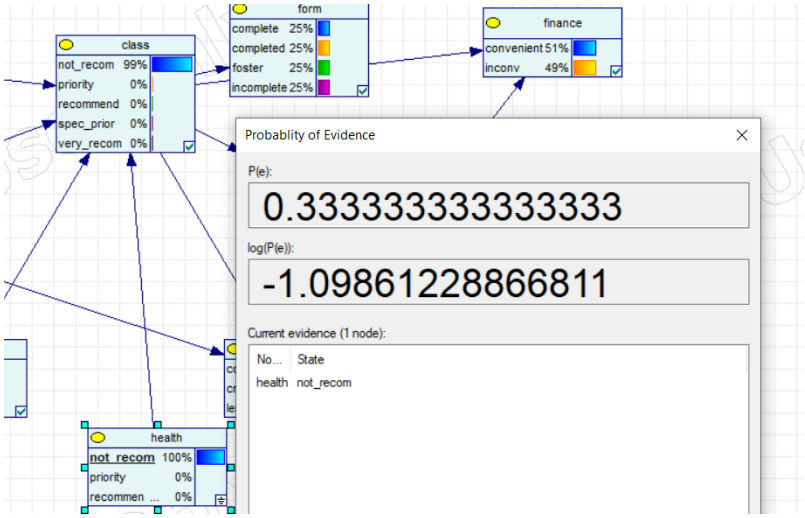
Analysis of the inference result: As a conclusion, we can say that the health conditions of a family, contributes in making the application prioritized or giving it the special priority acceptance.



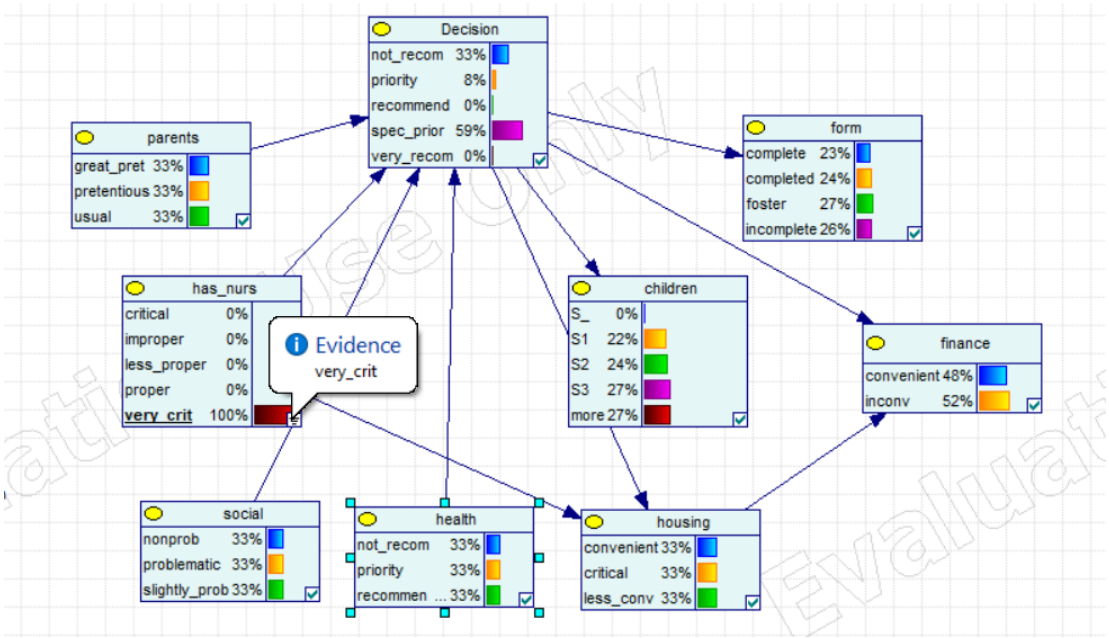
Let's see now if we set it to "not recommended" in health input, we will visualize a very strong impact on the decision of application acceptance.



With a probability of evidence equals to :



⇨ **Setting evidence:** Has_nurs to “very critical”



Analysis of the inference result: As a conclusion, we can say that the has_nursery input impacts the acceptance probability of the application.

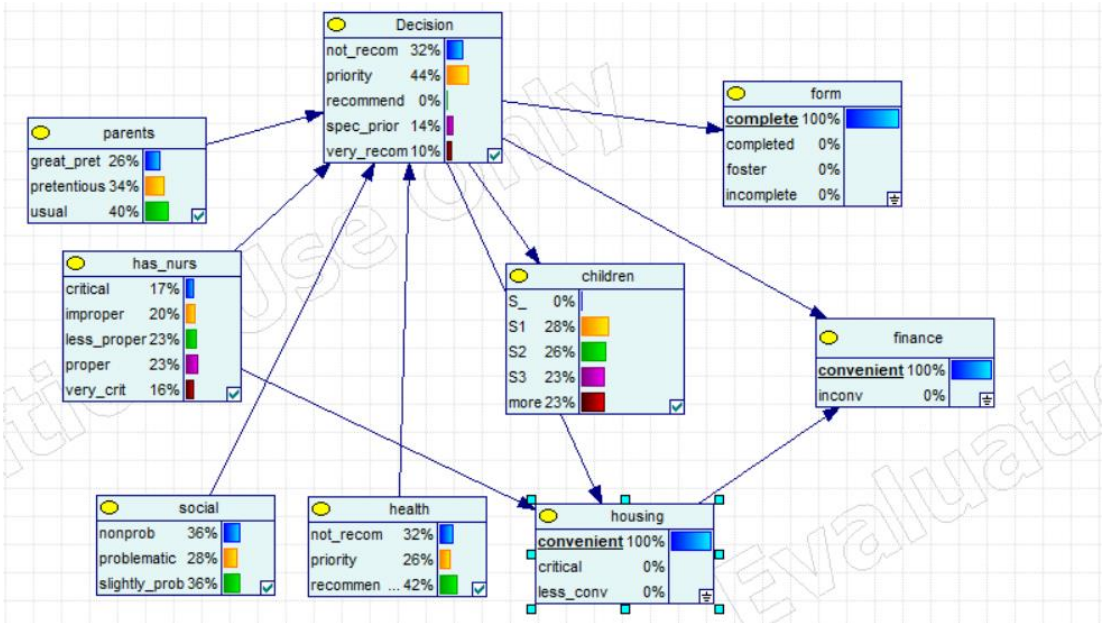
Now, let’s see if the family structure and financial standing will affect the application decision as stated at

first, we set:

- ✓ Form of the family: complete family.
- ✓ Housing conditions: convenient.
- ✓ Financial standing of the family: convenient.

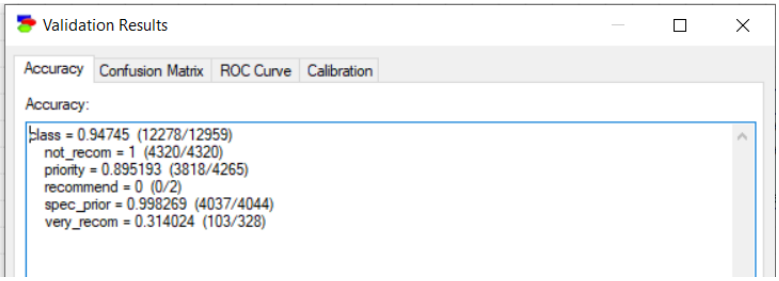
After updating beliefs, we see that is it recommended to accept the application of this family in view of its

structure and financial standing as shown in the next page.



✓ **Validation test**

Finally, we perform a validation test to our network.



V. Discussions:

In this work, we've began by learning graph and parameters for the Bayesian network from the chosen dataset in order to learn the existing casual relationships among variables. After this, we preformed inference (both exact and approximate) to produces the probability distribution of variables given one or more other variables.

Furthermore, we saw that applying inference manually with setting evidence, updating belief, and checking the updated probability gives much more observable results and explanation either for only one variable or by combining multiple factors.

Finally, we added this table below in order to show how the results will be presented to the families as a list of children's names arranged according to priorities.

Table 4: Applications, ranked according to the acceptance priority

Name	Evaluation attributes (leave values)								Evaluation result (formula)	Priority class (rules)
	K1	K2	K3	K4	K5	K6	K7	K8		
1.Child One	1	4	1	4	3	2	3	2	(4.48)	5
2.Child Two	2	3	3	2	3	2	3	1	(4.46)	5
3.Child Three	1	3	3	3	1	2	2	2	(4.08)	4
4.Child Four	2	3	3	1	3	2	2	1	(3.81)	4
.....										
27.Child Twenty-seven	2	3	1	1	3	1	2	1	(3.11)	4
.....										
38.Child Thirty-eight	2	3	1	1	1	1	1	1	(2.61)	4
39.Child Thirty-nine	2	3	1	1	1	1	1	1	(2.61)	4
40.Child Forty	1	3	1	3	1	1	1	1	(3.11)	3
41.Child Forty-one	2	1	3	1	2	1	1	1	(2.95)	3

Conclusion

To sum up, we must state that this process allowed to explain the evaluation results to the parents since once they were acquainted with the list that rank the applications according to acceptance priority, they could compare their children's positions with the others without attacking the schools boards and start questioning the reliability of the acceptance committee .

According to the main article several nursery schools and other professionals involved in the problem have adopted the approach and it was observed that the amount of work became smaller and the consistency of decisions higher. As a consequence, the number of conflict situations between parents and nursery schools decreased.

Acknowledgement : This Project was mainly based on the following research (Chapter 10: An application for admission in public school systems , Marko Bohanec, Vladislav Rajkovic, Manuel Olave)

References

[1] https://arxiv.org/ftp/arxiv/papers/1301/1301.6684.pdf

[2] http://kt.ijs.si/MarkoBohanec/pub/icml97.pdf

[3] https://archive.ics.uci.edu/ml/datasets/nursery

[4] http://www.cs.ru.nl/~peterl/teaching/CI/learn4.pdf

[5] https://support.bayesfusion.com/docs/GeNIe/hello.html

[6] https://bi.snu.ac.kr/Courses/4ai16s/slides/BayesianNetworks_practice1.pdf

[7] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.2022&rep=rep1&type=pdf