Wroclaw University of Science and Technology

Artificial Intelligence and Machine Learning

Politechnika
Wrocławska

Bayesian Belief Networks Project

Ines de Oliveira Soares, 256652

Problem Selection

After searching a lot of datasets on the internet, I was able to find a website with lots of datasets of real problems (https://www.kaggle.com/). I studied some datasets and I ended up choosing the Suicide Dataset to analyse. As this dataset was very big, I limited to the suicides that occurred in 2016.

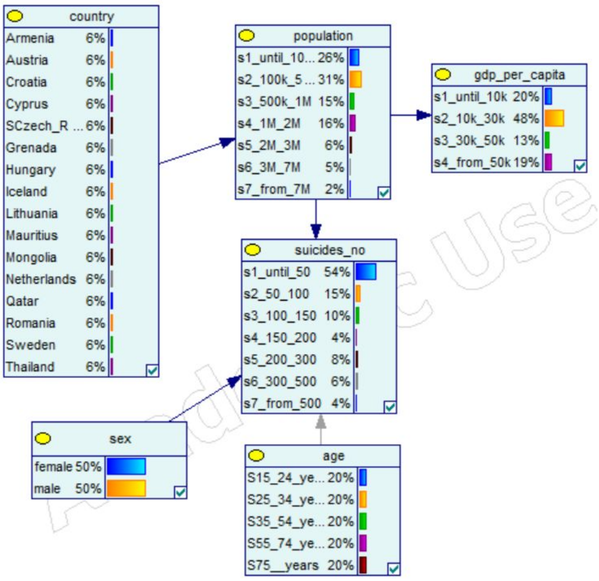
As explained in the presentation of the project, for this assignment we don't want a network very large. Because it would be too difficult to accurately calculate and verify all probabilities and correctness of the model in full detail, due to the fact that it requires an extensive investigation. That is why I am only working with six variables. The data set had 160 cases and the following variables: GDP-per-capita, Country, Population, Sex, Age and Number of Suicides.

Building and verifying a network

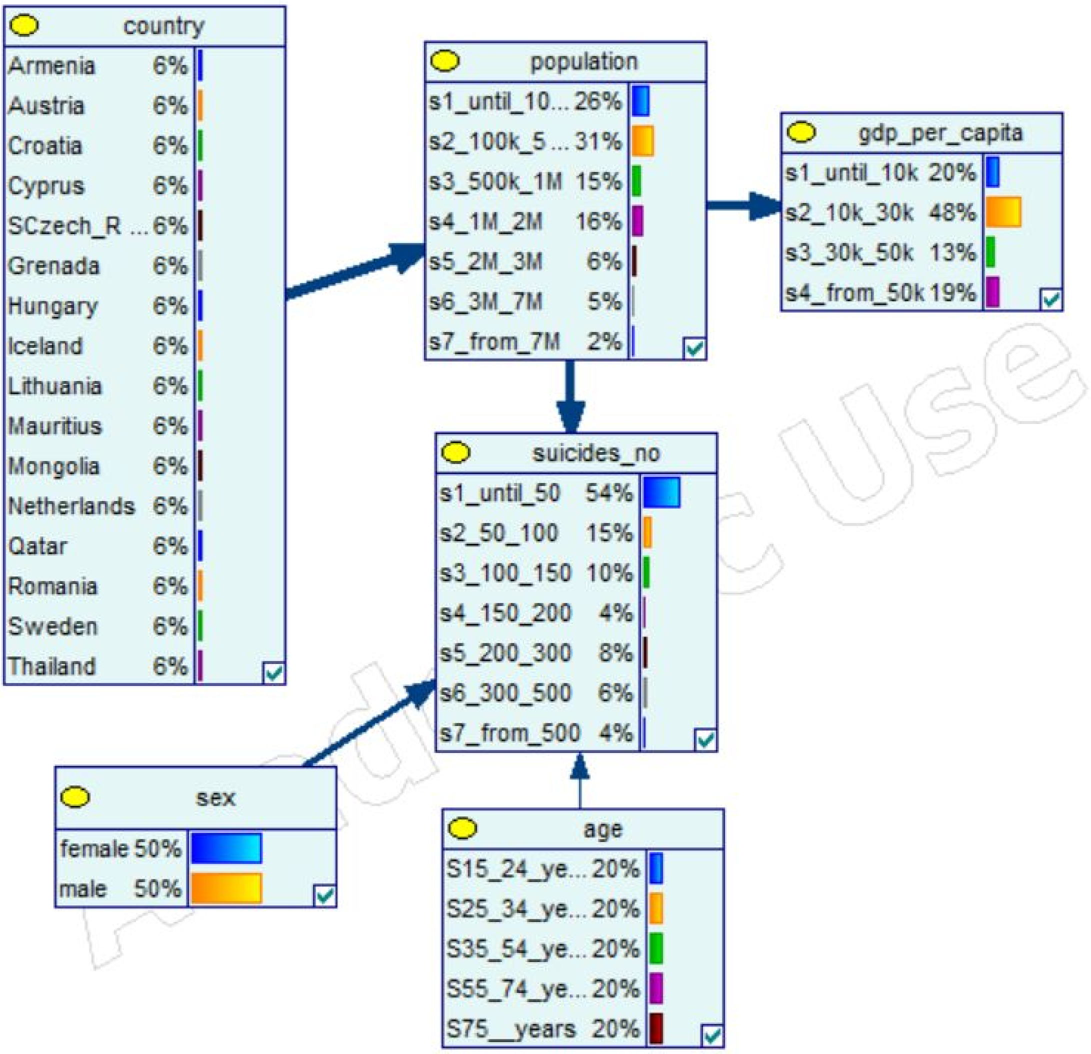
To build the Belief Network Problem I used the GeNie Academic software. But before uploading the dataset, it was necessary to modify it a bit because to generate the network, the software didn't allow continuous data, so variables like GDP-per-capita, Population, Age and Number of Suicides needed to be divided into classes.

The age was divided into five parts: ages between 15 and 24, then 25 and 34, 35 and 54, 55 and 74 and more than 75 years old. The number of suicides was divided into seven parts: less than 50 suicides, between 50 and 100, 100 and 150, 150 and 200, 200 and 300, 300 and 500 and more than 500. The population was divided also into seven parts: less than 100k, between 100k and 500k, 500k and 1M, 1M and 2M, 2M and 3M, 3M and 7M and more than 7M. And, finally, the gdp-per-capita was divided into four classes: less than 10k, between 10k and 30k, 30k and 50k and more than 50k dollars.

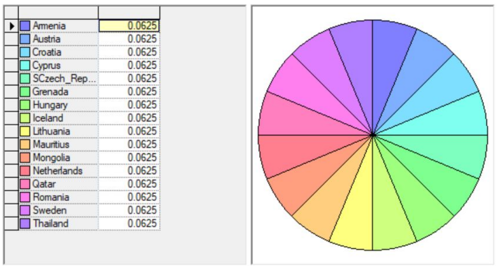
After modifying and verifying everything, I was able to generate the network and the result was the following:



In a network there are nodes that influence more than others. So it is possible to evaluate the strength of a node by analyzing the influence in the network if we remove it. In the graph below it is shown the strengths of the relations, this is represented by the thickness of the arcs. This way we can verify that the stronger relations are between country and population, population and gdp-per-capita, population and number of suicides.

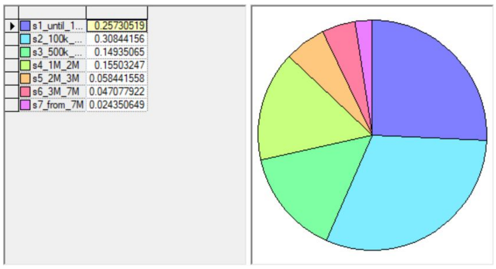


In the following graphs it is shown the properties of each nodes, more specifically, the probabilities of the variables of each node.



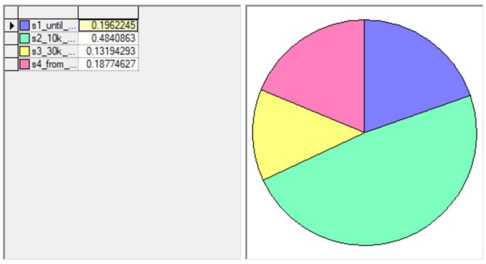
Armenia	0.0625
Austria	0.0625
Croatia	0.0625
Cyprus	0.0625
SCzech_Rep...	0.0625
Grenada	0.0625
Hungary	0.0625
Iceland	0.0625
Lithuania	0.0625
Mauritius	0.0625
Mongolia	0.0625
Netherlands	0.0625
Qatar	0.0625
Romania	0.0625
Sweden	0.0625
Thailand	0.0625

Country



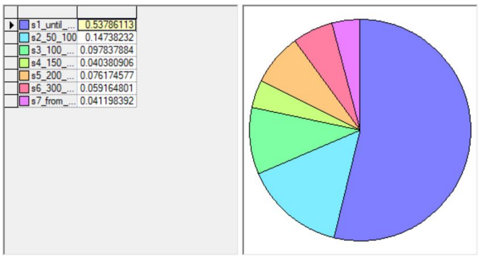
s1_until_1...	0.25730519
s2_100k_...	0.30844156
s3_500k_...	0.14935065
s4_1M_2M	0.15503247
s5_2M_3M	0.058441558
s6_3M_7M	0.047077922
s7_from_7M	0.024350649

Population



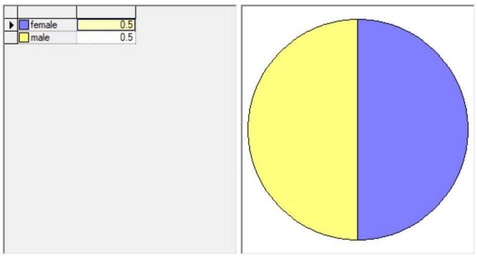
s1_until_...	0.1962245
s2_10k_...	0.4840863
s3_30k_...	0.13194293
s4_from_...	0.18774627

GDP-per-capita



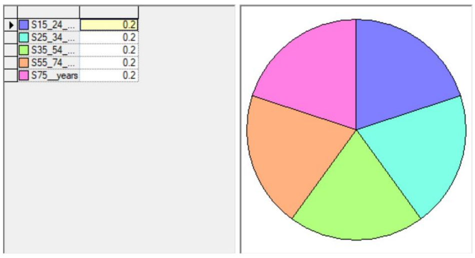
s1_until_...	0.53786113
s2_50_100	0.14738232
s3_100_...	0.097837884
s4_150_...	0.040380906
s5_200_...	0.076174577
s6_300_...	0.059164801
s7_from_...	0.041198392

Number of suicides



female	0.5
male	0.5

Sex



S15_24_...	0.2
S25_34_...	0.2
S35_54_...	0.2
S55_74_...	0.2
S75__years	0.2

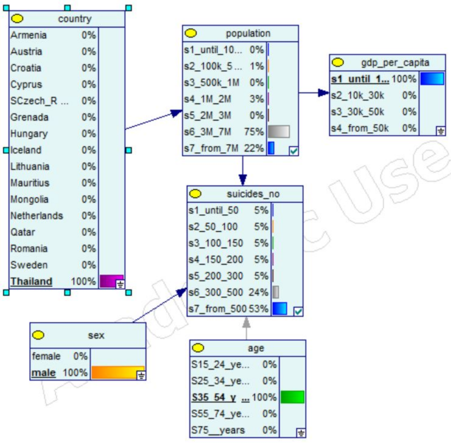
Age

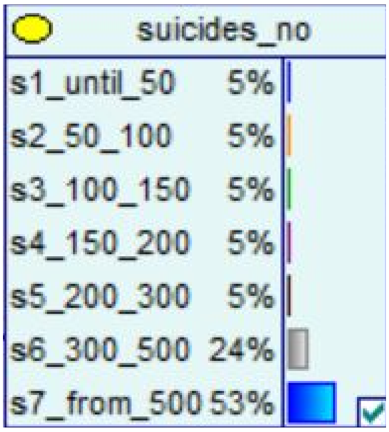
Computing probabilities with the network

After properly verifying the model, the objective was to compute probabilities with the network: posing queries for which answers are difficult to predict based on the data alone. So below it is presented some queries that show the results to these difficult answers.

Query 1

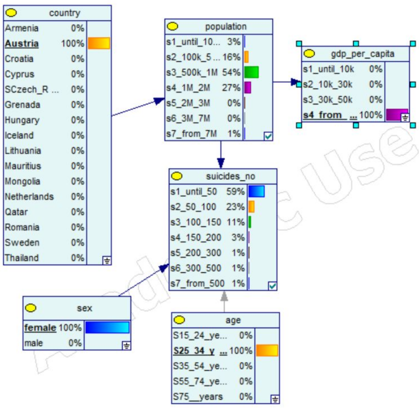
In this query, the objective is to show the results for men with age between 35 and 54 years old who live in Thailand and which gdp-per-capita is less than 10k dollars. And it is possible to verify that the number of suicides in these conditions is very high.

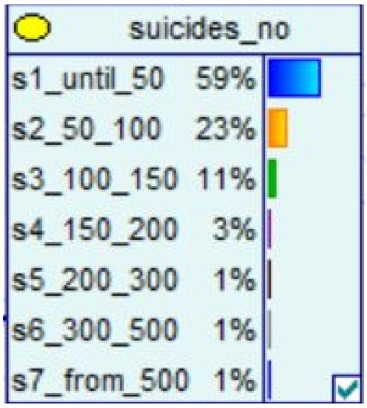




Query 2

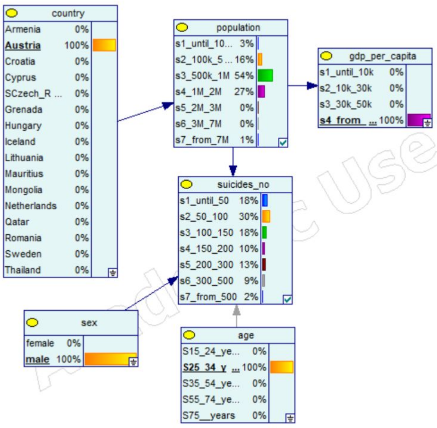
In this query, the objective is to show the results of women with age between 25 and 34 years old who live in Austria and which gdp-per-capita is higher than 50k dollars. And it is possible to verify that the number of suicides in these conditions is low.

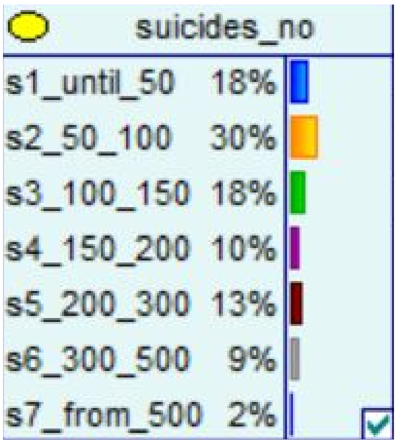




Query 3

In this query, the objective is to analyse the same condition as in the previous query (query 2) but instead of being women, this time it is men. And we can verify that the number of suicides increased, this way we can conclude that in these conditions men are more likely to commit suicide than women.

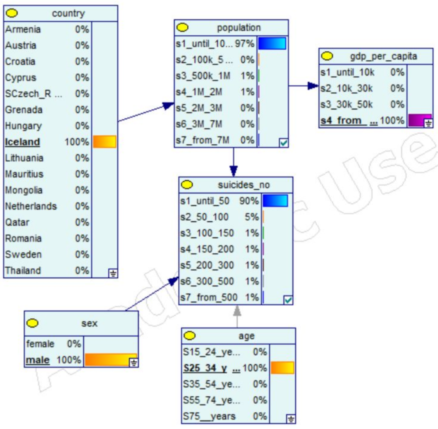


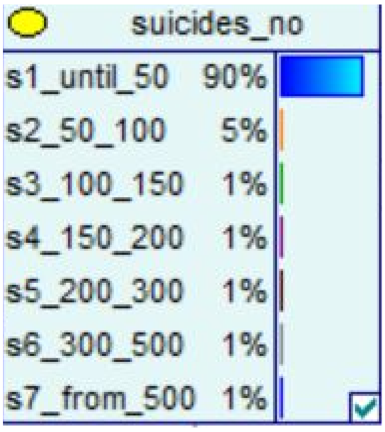


suicides_no	
s1_until_50	18%
s2_50_100	30%
s3_100_150	18%
s4_150_200	10%
s5_200_300	13%
s6_300_500	9%
s7_from_500	2%

Query 4

In this query, the objective is to analyse the same condition as in the previous query (query 3) but instead of being men from Austria it is man from Iceland. And we can verify that the number of suicides decreased a lot, this way we can conclude that in these conditions men in Iceland are less likely to commit suicide than men in Austria.





suicides_no	
s1_until_50	90%
s2_50_100	5%
s3_100_150	1%
s4_150_200	1%
s5_200_300	1%
s6_300_500	1%
s7_from_500	1%

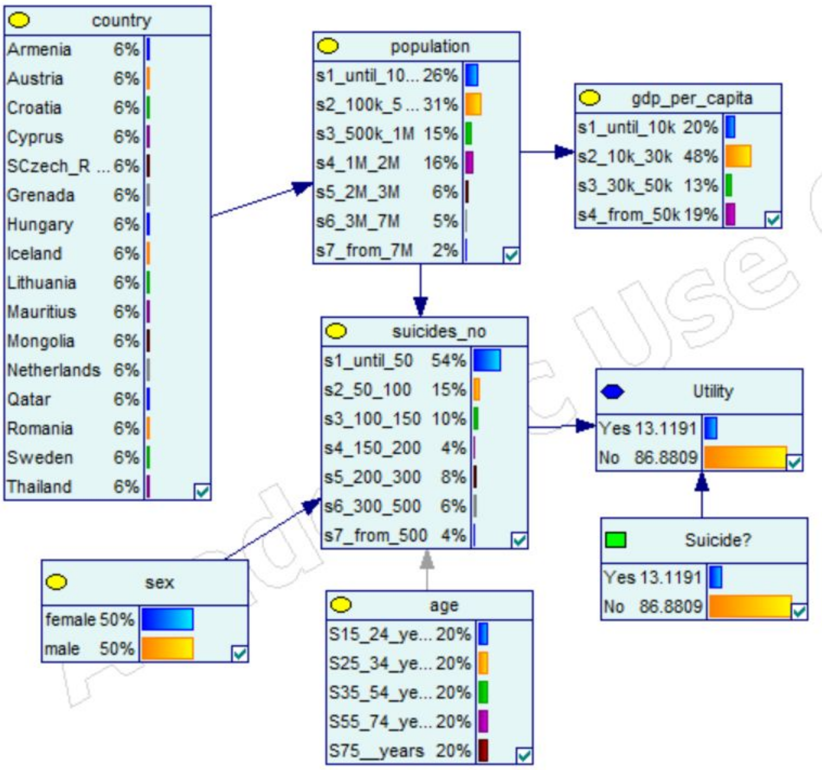
With this queries we can verify that the results than we obtain depend on the country, sex, age and gdp-per-capita that we choose. If each node was divided into more parts, it could be possible to do a more detailed search and obtain results more accurate. But here the objective is to have a general idea of the influence that each node has in the result.

Making decisions

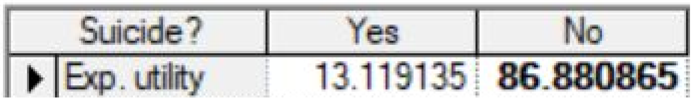
In this part the objective is to use the probabilities obtained from the belief network to make decisions. For this it is necessary to define the decision agent (number of suicides) and also define the utility function. The utility values that I defined for each possibility are shown in the table below.

Number of Suicides	Yes	No
less than 50	0	100
between 50 and 100	10	90
between 100 and 150	20	80
between 150 and 200	30	70
between 200 and 300	40	60
between 300 and 500	50	50
more than 500	60	40

In the following image it is shown the resulting network with the addition of the utility and decision nodes.

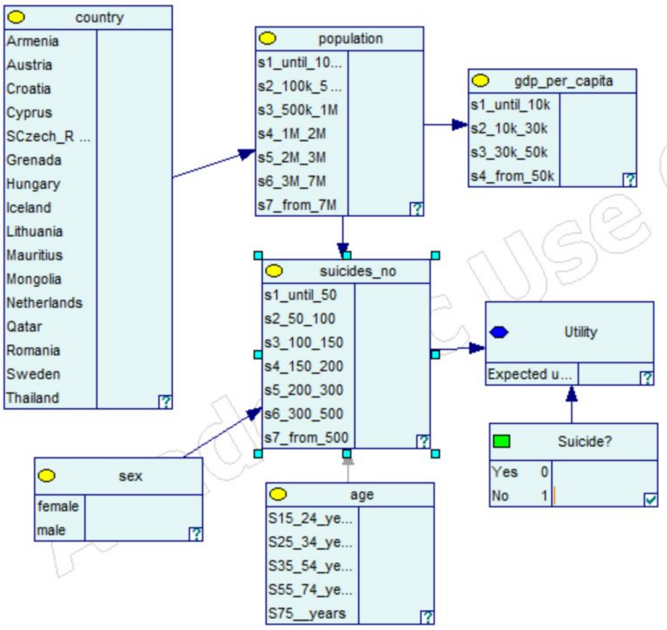


The expected utility of the situations resulting from the current information is shown in the following table.



Suicide?	Yes	No
Exp. utility	13.119135	**86.880865**

Using the Find Best Policy algorithm we obtain the following result.



With all this study it is possible to conclude that the optimal action for the agent is not to commit suicide.

Computing the value of perfect information

The value of information theory allows to make decisions about which information the agent should collect. The importance of information depends on two factors:

- whether various possible outcomes will significantly affect the decision
- the probabilities of different outcomes

In order to compute the value of perfect information I chose the node of population to be the chance node and obtain the following expected utilities for different policies:

population	**Yes**	**No**
until_100k	1.5513834	98.448617
100k_500k	7.6236264	92.376374
500k_1M	10.77381	89.22619
1M_2M	23.736264	76.263736
2M_3M	39	61
3M_7M	37.5	62.5
from_7M	42.5	57.5

So we can conclude that the maximum expected utilities are:

population	**until_100k**	**100k_500k**	**500k_1M**	**1M_2M**	**2M_3M**	**3M_7M**	**from_7M**
MEU	98.448617	92.376374	89.22619	76.263736	61	62.5	57.5

As we know from before, the maximum utility of *Suicide?* is 86.880865, corresponding to the action of not commiting suicide. And also we know the probabilities of the variables of population:

population	**until_100k**	**100k_500k**	**500k_1M**	**1M_2M**	**2M_3M**	**3M_7M**	**from_7M**
MEU	26%	31%	15%	16%	6%	5%	2%

So now it is possible to compute the value of perfect information:

$$\begin{aligned}\text{VPI(population)} = {} & P(\text{until\_100k}) * \text{MEU}(\text{until\_100k}) + P(\text{100k\_500k}) * \text{MEU}(\text{100k\_500k}) + \\ & + \; P(\text{500k\_1M}) * \text{MEU}(\text{500k\_1M}) + P(\text{1M\_2M}) * \text{MEU}(\text{1M\_2M}) + \\ & + \; P(\text{2M\_3M}) * \text{MEU}(\text{2M\_3M}) + P(\text{3M\_7M}) * \text{MEU}(\text{3M\_7M}) + \\ & + \; P(\text{from\_7M}) * \text{MEU}(\text{from\_7M}) - \text{MEU}(\text{no})\end{aligned}$$

$$\begin{aligned}\text{VPI(population)} = {} & 0.26 * 98.448617 + 0.31 * 92.376374 + 0.15 * 89.22619 + \\ & + \; 0.16 * 76.263736 + 0.06 * 61 + 0.05 * 62.5 + 0.02 * 57.5 - 86.880865\end{aligned}$$

$$\text{VPI(population)} = 0.87357762$$

For the utility distribution defined the value of population is approximately 0.87, expressed in utility units. This means that if we knew this information, we could improve our decisions by 0.87. So if the cost of acquiring this information is low, it could be profitable to know it. But if the cost is too high it is not worth it because it won’t change much our decision. For example, if we compute the value of information of the node *sex*, the result is zero. This means that whether we acquire or not the information, it won’t change our decision.