Artificial Intelligence DT8012 - Laboration 3

Jonas Fockstedt

January 23, 2020

1 Introduction

The purpose of this lab was to get a better understanding of how Bayesian Networks (BN) operate. It consisted of using the software Bayes Server[1], where it would be given one of the available data sets from the laboration supervisor and do Bayesian calculations on the variables. The idea of BN's is to calculate the probability of some event happening, given that another event has happened. This by using Bayes' Theorem, as show in Equation 1. The equation is widely used within research and Artificial Intelligence, where it can be useful to update one's belief on something given that something else has happen prior. This is especially useful for machine learning, where a machine constantly has to update its beliefs of certain events.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Bayes Server has a number of different functionalities and use areas, but for this laboration, the *Parameter learning*[2] and *Structure learning*[3] functions will be the ones of focus for this work. When running the *Parameter learning* function, it returns values of *Log likelihood*[4] and *Bayesian Information Criterion*[1] (BIC).

- *Log likelihood* - The probability that the given network is true, i.e. how likely that the given network has generated the given data. The lower the value, the less likely.
- *BIC* - A metric for defining how well the model fits the data set. This value is wished to be minimized as much as possible, meaning that the given network represent the data set the most.

2 Method

The were two given files, *data_artificial.xlsx* and *data_real.xlsx*, where the *data_artificial.xlsx* file was an artificial one with 700 observations and the *data_real.xlsx* file a real dataset with 1657 observations. The task was to first used the *data_artificial.xlsx* file to generate nodes and draw random connections between them. This would be done three times for different network where each network would be using the *Parameter learning* function to determine the log likelihood and the BIC of the given network. Finally, the network would itself learn the structure between the nodes by using the *Structure learning* function in Bayes Server. This procedure would then be repeated for the *data_real.xlsx* file, where additional probability calculations were made. Task 1 covers the *data_artificial.xlsx* file and task 2 covers the *data_real.xlsx* file.

3 Results

The results will be mostly displayed by pictures. For every network, there will be a figure showing the network as a whole, followed up by a figure describing the log likelihood and BIC for the given network.

[1] https://www.bayesserver.com/
[2] https://www.bayesserver.com/docs/learning/parameter-learning
[3] https://www.bayesserver.com/docs/learning/structural-learning
[4] https://www.bayesserver.com/docs/queries/log-likelihood

3.1 Task 1

The first network, as shown in Figure 1, shows how the different nodes were set up between one another. When the *Parameter learning* function was done on these connections, the values of *Log likelihood* and *BIC* is shown in Figure 2.

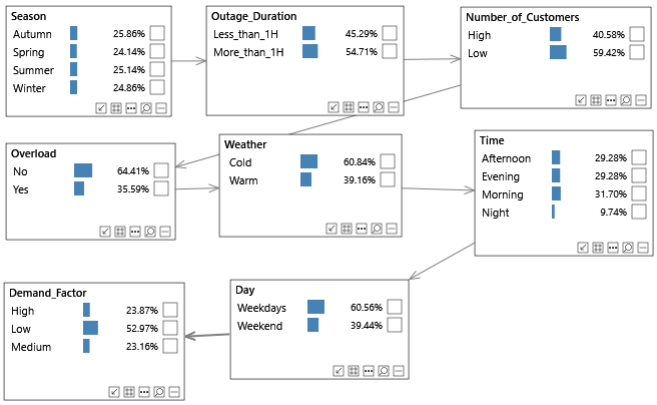

Figure 1: First network from the *data_artificial.xlsx* file.

	Created	Converged	Iteration Count	Log Likelihood	BIC
✕	1/20/2020 6:36:10 PM	☑	2	-4779.95435767668	9736.78788439954

Figure 2: *Log likelihood* and *BIC* from first network.

The second network, shown in Figure 3 displays a slightly different setup compared to the first network. Rather than the *Time* node would be connected to the *Day* node, and the *Day* node would be connected to the *Demand_Factor* node, it has switched order. Here, the *Time* node is connected to the *Demand_Factor* node, and the *Demand_Factor* in turn is connected to the *Day* node. As shown in Figure 4, this small change has lead to an increase in the *Log likelihood* by around 80, and the *BIC* factor is now 100 less than before.

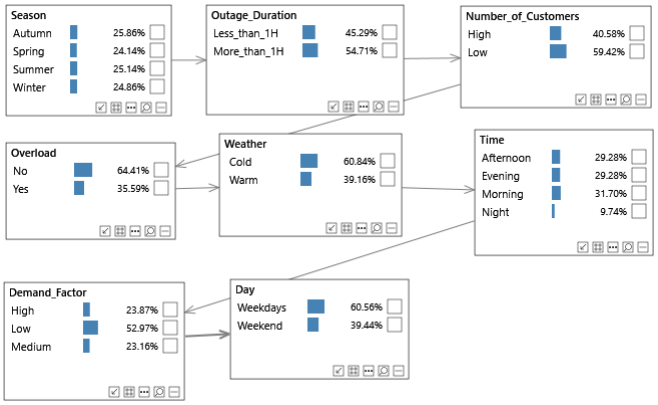

Figure 3: Second network from the *data_artificial.xlsx* file.

	Created	Converged	Iteration Count	Log Likelihood	BIC
✕	1/20/2020 7:45:36 PM	☑	2	-4719.45959623804	9635.45160252739

Figure 4: *Log likelihood* and *BIC* from second network.

In the third network, as shown in Figure 5, the network is now shown as the second network, but in reverse. A setup like this gives the output as shown in Figure 6. This gives the exact same values for both *Log*

likelihood and *BIC* as the second network got. The reasoning for this is probably that the calculations are done with respect to that the nodes are independent of one another, meaning that the reversed order does not affect the results.

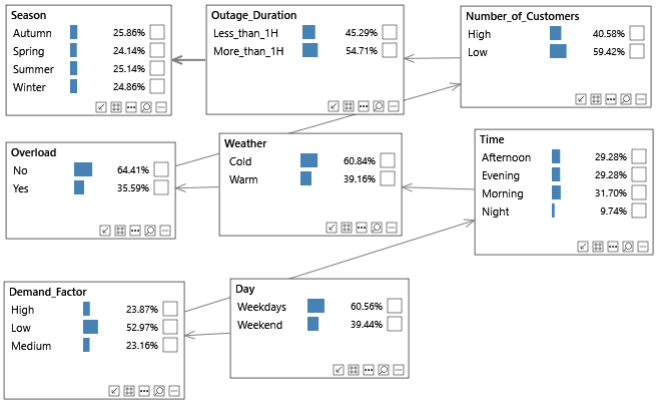

Figure 5: Third network from the *data_artificial.xlsx* file.

Created	Converged	Iteration Count	Log Likelihood	BIC
1/20/2020 7:48:53 PM	☑	2	-4719.45959623804	9635.45160252739

Figure 6: *Log likelihood* and *BIC* from third network.

Instead of constructing the links manually, the fourth network would be constructed by the *Structure learning* method in Bayes Server. The resulting network is shown in Figure 7, where the *Log likelihood* and *BIC* values are shown in Figure 8. Here, one can see that the *Log likelihood* value has gone up by just under 300, and the *BIC* value has gone down by around 500, compared to the second and third networks. These two improved factors indicate that this network is more optimal than the others, but it may not be the most optimal one.

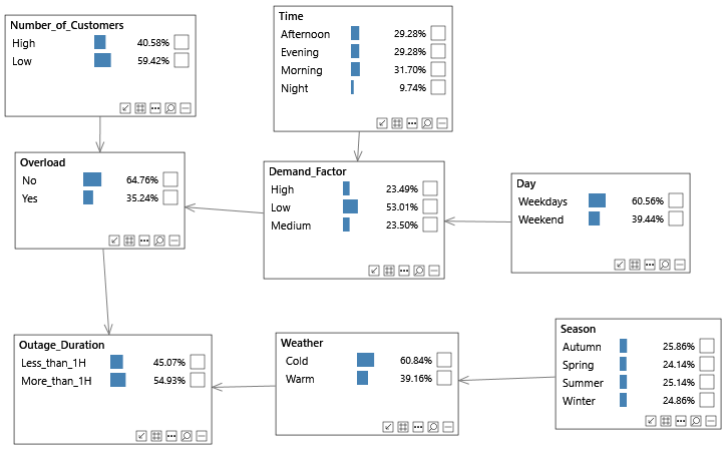

Figure 7: Fourth network from the *data_artificial.xlsx* file. This network was generated using the *Structure learning* method.

Created	Converged	Iteration Count	Log Likelihood	BIC
1/21/2020 11:04:22 AM	☑	2	-4468.78032075825	9186.50169424815

Figure 8: *Log likelihood* and *BIC* from fourth network.

4 Task 2

The first network based on the *data_real.xlsx* file is shown in Figure 9, where the corresponding *Log likelihood* and *BIC* values are shown in Figure 10. The connection between the nodes were set up manually.

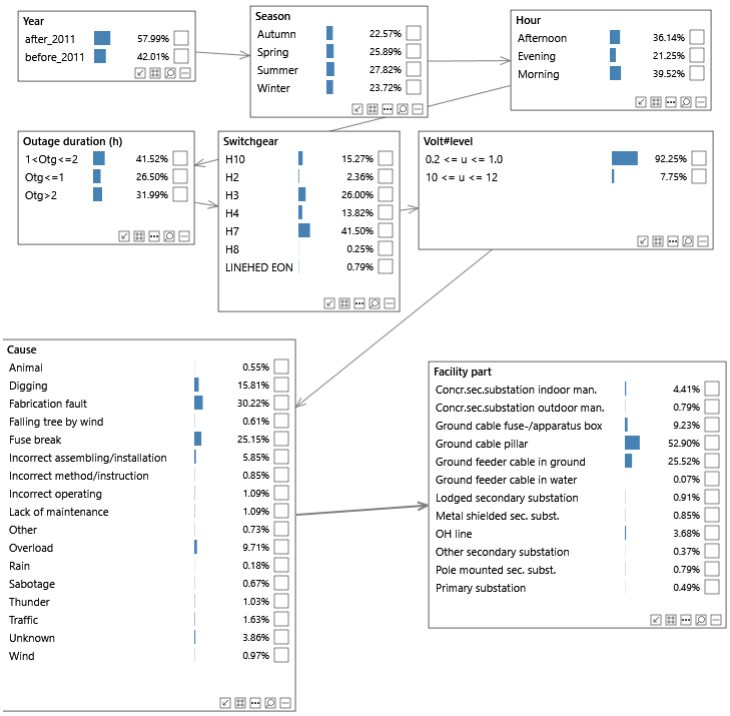

Figure 9: First network from the *data_real.xlsx* file.

Created	Converged	Iteration Count	Log Likelihood	BIC
1/21/2020 1:05:02 PM	☑	2	-14858.0437422157	31724.946533154

Figure 10: *Log likelihood* and *BIC* from first network.

The second network is shown in Figure 11, with the *Log likelihood* and *BIC* values shown in Figure 12. This network differs from the first network by letting the *Volt#level* node connect with the *Facility part* node, instead of letting the *Volt#level* node connect with the *Cause* node. The *Facility part* is then connected to the *Cause* node. This network has generated a higher *Log likelihood* value than the previous network, and also a lower *BIC* value. The new values indicates that this network presents the data better than the previous network.

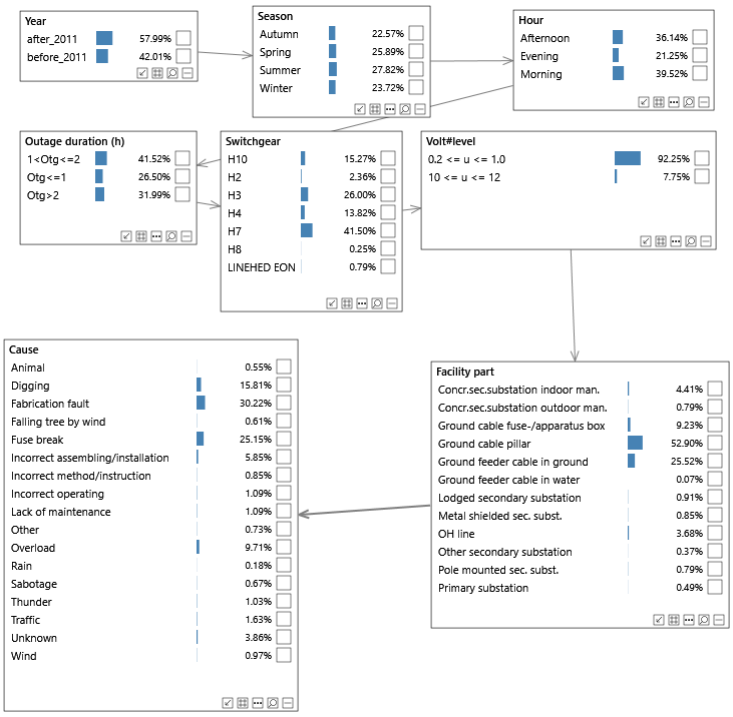

Figure 11: Second network from the *data_real.xlsx* file.

	Created	Converged	Iteration Count	Log Likelihood	BIC
✕	1/21/2020 1:07:05 PM	☑	2	-14765.2184312727	31502.2320911809

Figure 12: *Log likelihood* and *BIC* from second network.

For the third network (shown in Figure 13), The nodes are now connected in reversed order compared to the previous network. The *Log likelihood* and *BIC* values are shown in Figure 14, which gives the exact same values. This is due to the same reasoning as was mentioned for the second and third network in subsection 3.1, where it does not matter in which directions the nodes are connected when calculating the *Log likelihood* and *BIC* values.

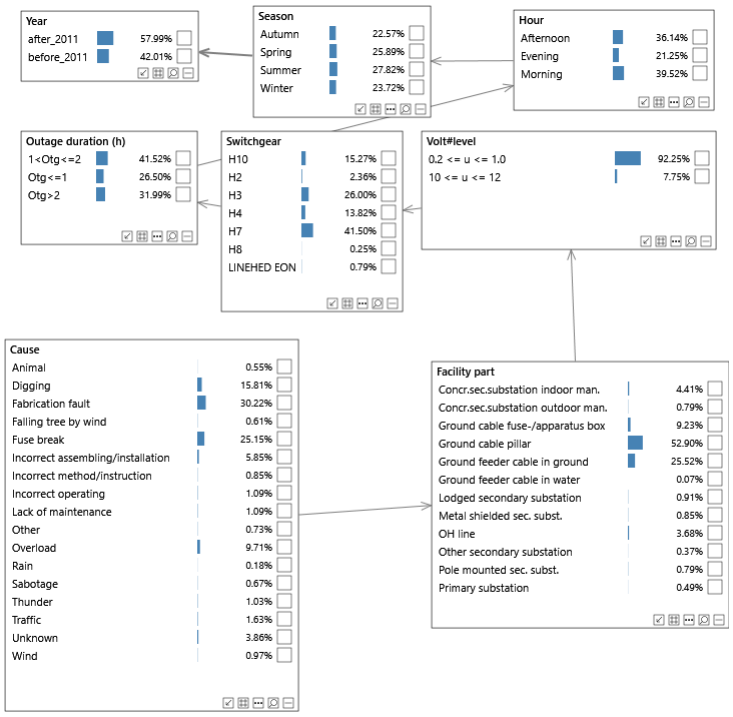

Figure 13: Third network from the *data_real.xlsx* file.

Created	Converged	Iteration Count	Log Likelihood	BIC
1/21/2020 1:10:37 PM	☑	2	-14765.2184312727	31502.2320911809

Figure 14: *Log likelihood* and *BIC* from second network.

The fourth network, shown in Figure 15, was constructed by using the *Structure learning* functionality to learn the relations between the data set. Interestingly enough, it detected that the nodes *Volt#level* and *Hour* does not matter, hence the lack of arrows to and from these nodes. This network generated a best this far *Log likelihood* value of -14090, but the biggest *BIC* value among the networks, at 69959, as shown in Figure 16. This may be due to that the two nodes, *Volt#level* and *Hour* are now not linked with any other node in the network.

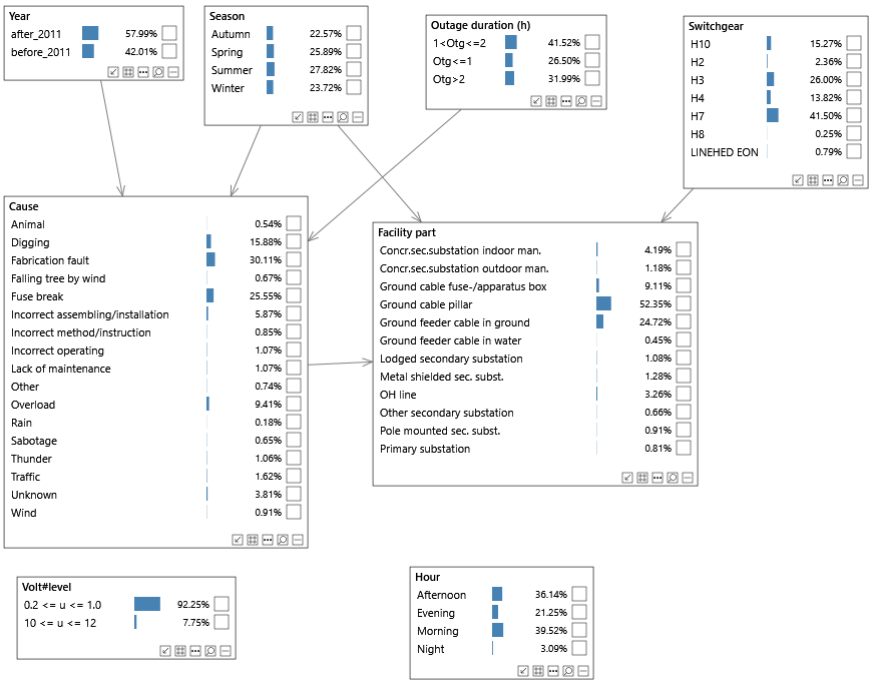

Figure 15: Fourth network from the *data_real.xlsx* file.

Created	Converged	Iteration Count	Log Likelihood	BIC
1/21/2020 1:13:11 PM	☑	2	-14090.5074716421	69959.3529455002

Figure 16: *Log likelihood* and *BIC* from second network.

4.1 Task 2b

Expression	**Value** (%)
P(Cause=Animal|Season=Autumn)	0.81
P(Season=Autumn|Cause=Animal)	33.64
P(Season=Summer|Cause=Thunder)	75.33
P(Outage duration=Otg $\leq$1|Facility part = Ground cable pillar)	29.85
P(Facility part=Ground cable pillar|Switchgear=H7, Cause=Fuse break)	81.79
P(Facility part=Ground feeder cable in ground|Cause=(Digging, Fabrication fault), Switchgear=H7, Season=Summer)	49.61
P(Facility part=Ground feeder cable in ground|Cause=¬(Digging, Fabrication fault), Switchgear=H7, Season=Summer)	27.08
P(Cause=Digging|Facility part = OH line, Switchgear=H7)	0
P(Facility part = Ground cable pillar|Outage duration=Otg >2)	45.21
P(Cause=Unknown|Year=before2011) / P(Cause=Unknown|Year=after2011)	2.46

Table 1: Probabilities of different types of scenarios from the fourth network.

4.2 Task 2c

Expression	**Value** (%)
P(Cause=Animal|¬Season=Autumn)	0.46
P(Season=Autumn|¬Cause=Animal)	0
P(Season=Summer|Cause=(Thunder, Rain))	100
P(Outage duration=Otg >2|Facility part = Ground cable pillar)	27.62
P(Facility part=Ground cable pillar|Switchgear=(H3,H7), Cause=Fuse break)	80.31
P(Facility part=Ground feeder cable in ground|Cause=(Digging, Fabrication fault), Switchgear=H7, Season=¬Summer)	36.39
P(Cause=Digging|Facility part = (OH line, Other secondary substation), Switchgear=H7)	0
P(Facility part = Ground cable pillar|Outage duration=1 < Otg <=2)	53.63
P(Season=Autumn|Cause=Lack of maintenance)	11.23
P(Facility part=Ground cable pillar|Switchgear=H7, Cause=Fuse break, Year=after_2011)	82.48

Table 2: Additional probabilities of different scenarios from the fourth network.

4.3 Task 2d

With respect to Equation 1, if one would want to calculate the probability of a child node, one must also know the conditional probability of its parent node, given that the child node has happened. Scenario - grass on a lawn can get wet by both the events of rain and that the sprinkler is on. If it rains, the probability of wet grass is increased, the same applies for if the sprinkler would be on. If, however, it is known that the grass is wet, then the probability has increased for that it has rained, or that the sprinkler has been on. These kinds of connections between events is what Bayes' Theorem suggests. A couple of examples with this effect is shown in

Child	**Parent**	**Expression**	**Value** (%)
P(Cause=Rain)	P(Season=Summer)	P(Season=Summer)	27.82
P(Cause=Rain)	P(Season=Summer)	P(Season=Summer|Cause=Rain)	0.52
P(Faculty part=OH line)	P(Switchgear=H3)	P(Switchgear=H3)	26.00
P(Faculty part=OH line)	P(Switchgear=H3)	P(Switchgear=H3|Faculty part=OH line)	5.87
P(Cause=Animal)	P(Year=after_2011)	P(Year=after_2011)	57.99
P(Cause=Animal)	P(Year=after_2011)	P(Year=after_2011|Cause=Animal)	65.28

Table 3: Demonstration of how parent nodes can be affected by their child nodes.

4.4 Task 2e

The combination of attributes that maximises the attributes *Ground feeder cable in ground* and *Fabrication fault* are shown in Table 4

Attribute to maximize	**Attributes**	**Value** (%)
Ground feeder cable in ground	Season=Summer, Switchgear=H7, Cause=Other	99.94
Fabrication fault	Season=Spring, Year=after_2011, Outage duration=Otg>2, Facility part=Other secondary substantion, Switchgear=H3	99.85

Table 4: Attributes required in order to maximize *Ground feeder cable in ground* and *Fabrication fault*, respectively.

4.5 Task 2f

Table 5 and Table 6 displays high and low correlation between attributes, respectively.

Attribute	**Correlated to**	**Correlation** (%)
Ground feeder cable in ground	Digging	56.20
Primary substation	H8	8.18
Fuse break	1≤Otg	41.14

Table 5: Table of attributes with high correlation.

Attribute	**Correlated to**	**Correlation** (%)
Ground cable pillar	LINEHED EON	6.36
Wind	Spring	0.00
Rain	Otg>2	0.00

Table 6: Table of attributes with low correlation.

Most of the correlations between attributes does not come as a big surprise. For example, it is pretty reasonable that *Digging* may lead to faculty part that is *Ground feeder cable in ground*. Also, if the *Primary substation* has been struck with malfunction, it will most likely have to do with the *H8* switch gear, as it is 2 per cent units higher than the closest case, and almost 10 times as likely for the other cases. One can also interpret that, if it is an outage duration which is less than 1 hour, it will most likely be due to a fuse break and the problem could be resolved relatively quickly.

Whenever there is a *Ground cable pillar* malfunction, the probability of that any switch gear may have caused this is about 50%. However, if it is *LINEHED EON*, that probability drops down to 6.36%, which may be due that that switch gear has a better resistance against such malfunctions. The data sets also tells that, if it is *Wind*, then it is certainly not *Spring*, the same can be said about *Rain* and *Otg>2*. This may be due to that those particular days where the data was collected, it was never windy during the spring season. However, it seems a bit more logical that when there has been an outage of over 2 hours, it would most likely not be due to rain since a power grid should be able to withstand rainy weather. Unless it is a hurricane, which does not sound that likely to happen in Sweden.

References

[1] Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(2):199–203.