Early stage diabetes risk prediction

José Manuel Díaz Urraco

1. Introduction

Diabetes is a disease that can cause fatal consequences such as blindness, kidney failure, myocardial infarction, stroke, amputation of the lower limbs and, in the worst cases, death. In addition, its medical treatment is long and expensive. Therefore, getting an early diagnosis is crucial, not only to prevent health problems (and possible deaths) but also to save medical expenses.

Fortunately, various studies ([1], [2], [3] and many more) show that through the use of Data Mining techniques, tools can be developed to support professionals in achieving an early diagnosis of diabetes.

In this study, various analyzes have been carried out on a dataset that contains information about hospital patients who have symptoms related to diabetes. During these analyzes, different Bayesian networks have been learnt in order to explore relationships between the diabetes and the symptoms that patients present. Subsequently, exact and approximate inferences have been made to obtain knowledge of the trained networks. The software used to carry out this study has been GeNIe Modeler [4].

The analyzed dataset [5] has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet (Bangladesh) and approved by a doctor. The dataset has 520 instances, 16 predictor variables and the class variable, as shown in Table 1. Dataset variables.

No.	Variable name	Values	Type
1	Age	16-90	Numeric
2	Gender	Male, Female	Nominal
3	Polyuria	Yes, No	Nominal
4	Polydipsia	Yes, No	Nominal
5	sudden weight loss	Yes, No	Nominal
6	weakness	Yes, No	Nominal
7	Polyphagia	Yes, No	Nominal
8	Genital thrush	Yes, No	Nominal
9	visual blurring	Yes, No	Nominal
10	Itching	Yes, No	Nominal
11	Irritability	Yes, No	Nominal
12	delayed healing	Yes, No	Nominal
13	partial paresis	Yes, No	Nominal
14	muscle stiffness	Yes, No	Nominal
15	Alopecia	Yes, No	Nominal
16	Obesity	Yes, No	Nominal
17	class	Positive, Negative	Nominal

Table 1. Dataset variables

2. Problem description

Diabetes is a chronic and irreversible disease of the metabolism in which an excess of glucose or sugar is produced in the blood and in the urine. According to WHO [6], in 2019 an estimated 1.5 million deaths were directly caused by diabetes, and another 2.2 million deaths were attributable to high blood glucose in 2012. Furthermore, it also states that diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation.

The second main problem with this disease is its expensive medical treatment. According to ADA [7], the total costs of diagnosed diabetes have risen to $327 billion in 2017. This cost includes $237 billion in direct medical costs (hospital inpatient care, prescription medications to treat complications of diabetes, anti-diabetic agents and diabetes supplies, physician office visits) and $90 billion in reduced productivity (increased absenteeism, reduced productivity while at work for the employed population, reduced productivity for those not in the labour force, inability to work as a result of disease-related disability, lost productive capacity due to early mortality).

To reduce the impact of these problems, it has been proven that by identifying patients with pre-diabetes and initiating early interventions in lifestyle and/or pharmacological treatments, the progression of the disease can be delayed, or in some cases even prevented [8]. This is where Bayesian networks come into play, which are probabilistic graphical models that have been used to explore relationships between diabetes and various symptoms that a patient may present. Knowledge of these relationships can help professionals to find the symptoms that best explain or influence diabetes and, in this way, be able to detect diabetes in patients more quickly.

3. Methodology

The dataset was preprocessed first. Subsequently, various analyzes have been performed applying different algorithms in order to learn the **parameters** and the **structure** of three different Bayesian networks. In addition, **sensitivity analyzes** have been carried out (setting the *class* node as a target) to have an initial idea of which are the symptoms that can influence more on diabetes. Finally, exact and approximate **inferences** have been made to extract more precise knowledge of the learned networks. Figure 1. Study workflow shows the steps that have been followed during the study.

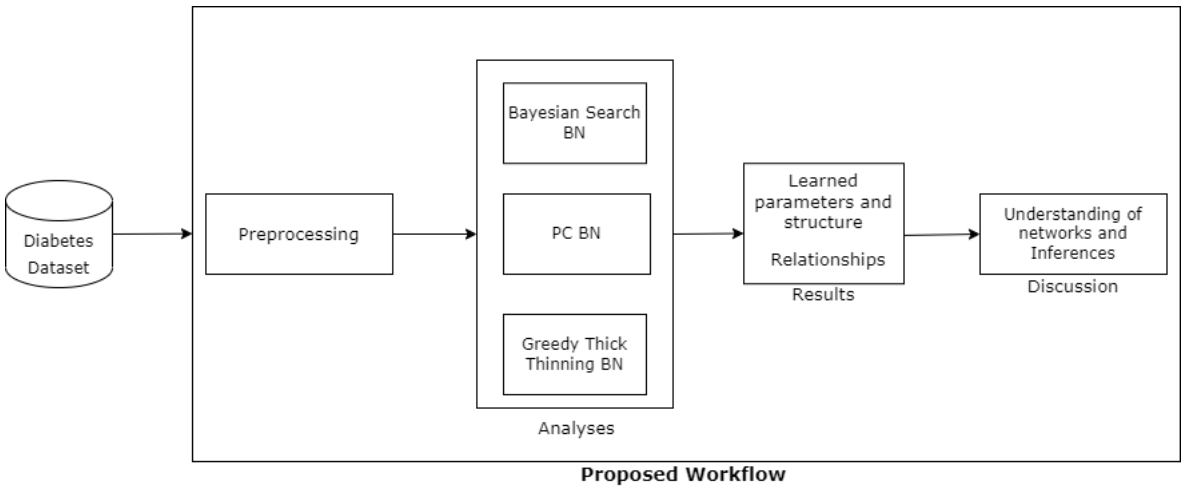

Figure 1. Study workflow

3.1. Preprocessing

First, the dataset has been visualized, the variables it presents can be observed in Table 1. Dataset variables. No missing values or outliers have been observed.

Also, during this phase, a series of observations were also made about the patients registered in the dataset:

1. Regarding **age**, there are 143 people between 18 and 39 years old, 281 people between 40 and 59 years old, 95 people between 60 and over, and only 1 person under 18 years old.
2. Regarding **gender**, 328 patients are male and 192 are female.
3. With regard to **symptoms**, the most frequent are: weakness (305 patients), polyuria (258 patients), itching (253 patients), delayed healing (239 patients), polyphagia (237 patients), polydipsia (233 patients) and visual blurring (233 patients as well). In turn, the least frequent are: obesity (88 patients), genital thrush (116 patients), irritability (126 patients) and alopecia (179 patients). Also note that there are 53 people (of the 520 in total) who do not have any symptoms and, of those 53 people, there are 6 who do have diabetes (class -> Positive).
4. Finally, regarding the **class** variable, there are 320 patients with diabetes and 200 without diabetes (class -> Negative).

None of the three selected algorithms to learn the structure of Bayesian networks supports numeric (continuous) variables mixed with nominal variables. Therefore, the *Age* variable **has been excluded from the networks** (it has not been considered necessary to discretize it since in previous studies it was determined that the patient's age was not an especially influencing factor in diabetes).

3.2. Analyses

As previously mentioned, three different Bayesian networks have been learned. In order to learn the **parameters**, the *EM algorithm* has been used in the three networks, it should be remembered that the objective of this algorithm is to find the parameters that characterize the model by maximum likelihood, that is, the higher the *EM Log Likelihood*, the more likely it is that data is generated by the estimated parameters. Regarding the **structure**, the following algorithms have been used: *Bayesian Search*, *PC* and *Greedy Thick Thinning*.

Once the networks have been trained, some of the **relationships** present in each of them have been analyzed (conditional dependencies, **marginal** independences and **conditional** independencies).

Subsequently, a **sensitivity analysis** was performed on each trained network. According to the official GeNIe documentation: "*Sensitivity analysis (Castillo et al., 1997) is technique that can help validate the probability parameters of a Bayesian network. This is done by investigating the effect of small changes in numerical parameters (ie, probabilities) on the output parameters (eg, posterior probabilities".* These analyzes have been used to obtain a **rough idea** of which are the nodes of the network that most influence diabetes (setting the *class* node as a target) and, in turn, which are the nodes that need to be paid some more attention when making inferences.

Finally, the *Clustering algorithm* has been applied to make **exact** inferences and the *Likelihood sampling algorithm* has been used to make **approximate** inferences. **Causal**, **diagnostic** and **inter-symptom** inferences have been made in the three networks considering the symptoms of the patients and whether they are diabetic or not (**disease**).

3.2.1. Bayesian Search network

In Figure 2. Bayesian Search parameters and EM score, the values of the parameters used to train the *Bayesian Search* network are observed. The *EM Log Likelihood* can also be observed, the greater it is, the better the network parameters have been learned.

Algorithm parameters:
Iterations: 20
Max parent count: 8
Sample size: 50
Link probability: 0.1
Prior link probability: 0.001
Seed: 0
Max search time: 0
No background knowledge

Best score in iteration 9: -4738.69
EM Log Likelihood: -4589.76

Figure 2. Bayesian Search parameters and EM score

3.2.2. PC network

In Figure 3. PC parameters and EM score, the values of the parameters used to train the *PC* network are observed. The *EM Log Likelihood* can also be observed, the greater it is, the better the network parameters have been learned.

Learning algorithm: PC
Algorithm parameters:
Max adjacency: 8
Significance: 0.05
Max search time: 0
No background knowledge

EM Log Likelihood: -4405.77

Figure 3. PC parameters and EM score

3.2.3. Greedy Thick Thinning network

In Figure 4. GTT parameters and EM score, the values of the parameters used to train the *Greedy Thick Thinning* network are observed. The *EM Log Likelihood* can also be observed, the greater it is, the better the network parameters have been learned.

Learning algorithm: Greedy ThickThinning
Algorithm parameters:
Max parent count: 8
No background knowledge

Score: -4666.04
EM Log Likelihood: -4058.3

Figure 4. GTT parameters and EM score

It can be seen that the most optimal network is apparently the one learned with the *Greedy Thick Thinning* algorithm since it presents the highest *EM Log Likelihood* (*-4058.3*).

4. Results

Next, some of the learned **parameters** (*marginal*, *conditional* and *joint* probability distributions) for the three networks will be shown together with their **structure**, found **relationships** and **independencies** (*conditional* and *marginal*).

4.1. Bayesian Search network

Regarding the **parameters**, in Figure 5. BSN: conditional probability distribution of class variable, the *conditional* probability distribution of the *class* variable given the *Polydipsia* variable can be

observed. Also, in Figure 6. BSN: joint probability distribution of class and Polydipsia variables, the *joint* probability distribution of the *class* and *Polydipsia* variables can be observed, together with the *marginal* distribution of *class* variable.

Polydipsia	No	Yes
Negative	0.66898955	0.034334764
Positive	0.33101045	0.96566524

Figure 5. BSN: conditional probability distribution of class variable

Joint probability distribution:

Polydipsia	No	Yes	Marginals
Negative	0.37595575	0.015039467	0.39099522
Positive	0.18601977	0.42298501	0.60900478

Figure 6. BSN: joint probability distribution of class and Polydipsia variables

In this case, the learned probabilities appear to be logical since not all patients in the dataset have polydipsia, therefore it makes sense that the joint probability of having diabetes and polydipsia together (*0.422*) is smaller than the probability of having only diabetes (*0.609*). In turn, once the patient already has polydipsia, it is much more likely that he also has diabetes (*0.965*).

Regarding the **structure** of the network, it can be seen in Figure 8. Bayesian Search network that the **parent** node of *class* is *Polydipsia* and the **child** nodes are *Gender* and *visual blurring*. Therefore, *class* conditionally depends on *Polydipsia*. In turn, *Gender* and *visual blurring* conditionally depend on *class*. Moreover, it can be seen that the parent node of *Polydipsia* is *Polyuria*, so *Polyuria* and *class* are **conditionally independent** given *Polydipsia*, that is, $\mathrm{P}(\mathrm{class}|\mathrm{Polyuria},\mathrm{Polydipsia}) = \mathrm{P}(\mathrm{class}|\mathrm{Polydipsia})$. Furthermore, the *Genital thrush*, *Irritability* and *Obesity* nodes are **marginally independent** of *class*, that is, $\mathrm{P}(\mathrm{class}) = \mathrm{P}(\mathrm{class}|\mathrm{Obesity})$, etc.

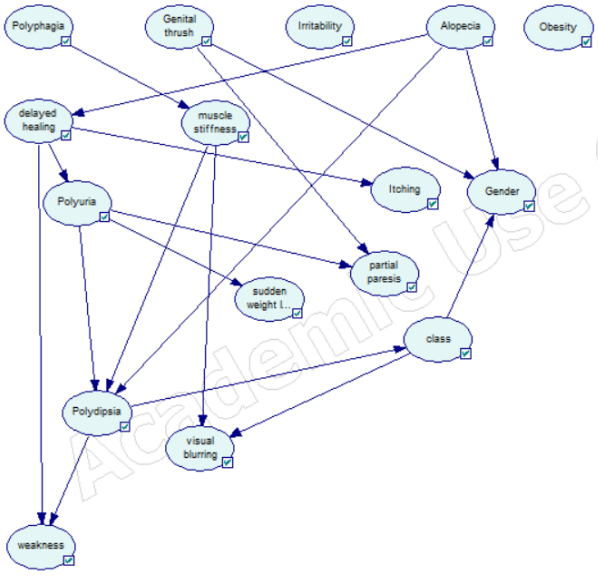

Figure 8. Bayesian Search network

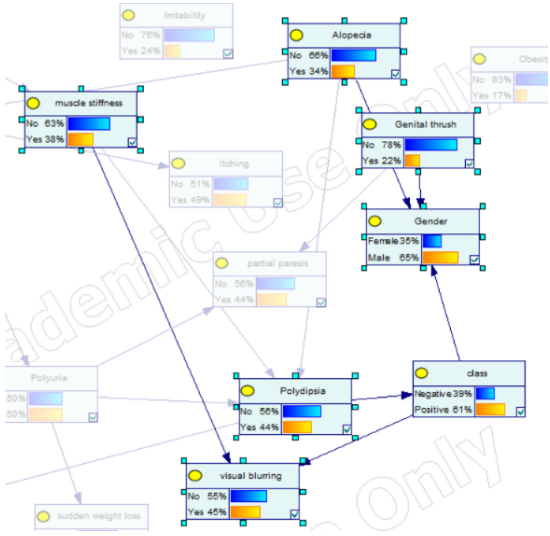

Figure 7. BSN: Markov blanket of class node

Figure 7. BSN: Markov blanket of class node shows the Markov blanket of the *class* node. Given the *visual blurring*, *Polydipsia*, *Gender*, *Genital thrush*, *Alopecia* and *muscle stiffness* nodes, the *class* node is **conditionally independent** from the rest of the network nodes (*sudden weight loss*, *partial paresis*, etc).

From a medical point of view, perhaps the structure of the network is not the most optimal. For example, polydipsia could be treated as a consequence of diabetes rather than as an antecedent (possible cause). Also, it does not make much sense to consider the gender of a patient as a consequent.

4.2. PC network

Regarding the **parameters**, in Figure 9. PC: conditional probability distribution of class variable, the *conditional* probability distribution of the *class* variable given the *Gender* and *Irritability* variables can be observed. Also, in Figure 10. PC: joint probability distribution of class, Gender and Irritability variables, the *joint* probability distribution of the *class*, *Gender* and *Irritability* variables can be observed, together with the *marginal* distribution of *class* variable.

Gender	Female		Male	
Irritability	No	Yes	No	Yes
Negative	0.125	0.020833333	0.664	0.19230769
Positive	0.875	0.97916667	0.336	0.80769231

Figure 9. PC: conditional probability distribution of class variable

Joint probability distribution:

Gender	Female		Male		Marginals
Irritability	No	Yes	No	Yes	
Negative	0.035837579	0.0017193...	0.31273847	0.030726453	0.38102188
Positive	0.25086305	0.080810758	0.1582532	0.1290511	0.61897812

Figure 10. PC: joint probability distribution of class, Gender and Irritability variables

As in the previous network, the computed probabilities make sense from a mathematical point of view. However, they do not seem logical from a medical point of view. For example, a female patient with irritability should not have such a high probability of having diabetes (*0.979*) since irritability is a common symptom of many other diseases, the same happens with male patients without irritability (*0.664* probability of not having diabetes). When comparing the conditional probabilities with the marginal ones, it is seen that the variation between the two is remarkably high considering that the irritability and the gender of a patient are quite generic factors.

Regarding the **structure** of the network, it can be seen in Figure 12. PC network that the parent nodes of *class* are *Gender* and *Irritability*. It is also observed that the child nodes of *class* are *Polyuria* and *Polydipsia*. Therefore, the *class* node conditionally depends on *Gender* and *Irritability*. In turn, the *Polyuria* and *Polydipsia* nodes are conditionally dependent on *class*. Moreover, it can be seen that *Polyphagia* is the child node of *Polyuria*. Therefore, *class* and *Polyphagia* are **conditionally independent** given *Polyuria*. In addition, it has been found that the *Obesity*, *visual blurring*, *sudden weight loss*, *Itching* and *weakness* nodes are **marginally independent** of the *class* node.

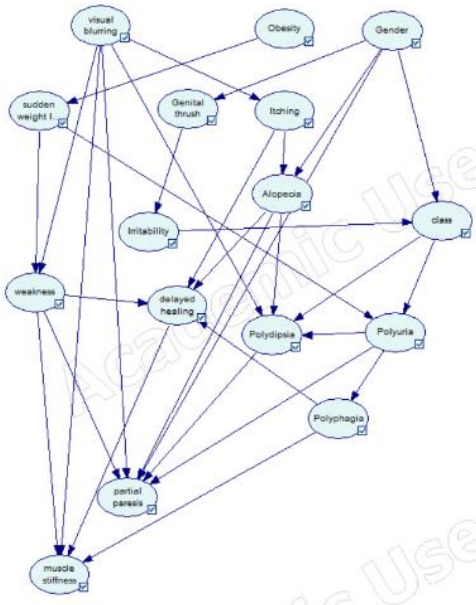

Figure 12. PC network

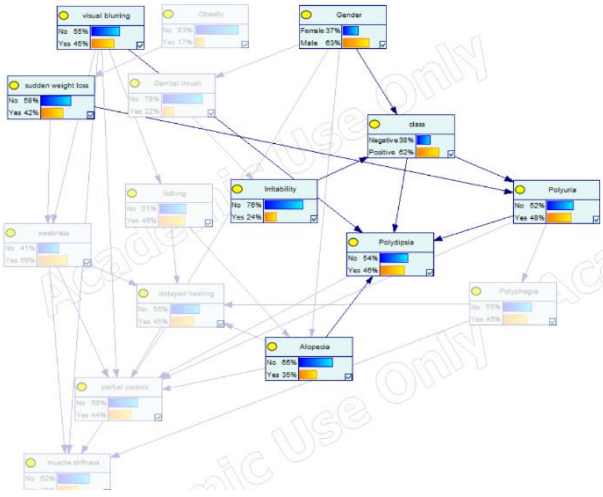

Figure 11. PC: Markov blanket of class node

Figure 11. PC: Markov blanket of class node shows the Markov blanket of the *class* node. Given the *visual blurring*, *Polydipsia*, *Gender*, *sudden weight loss*, *Alopecia*, *Irritability* and *Polyuria* nodes, the *class* node is **conditionally independent** from the rest of the network nodes (*weakness*, *partial paresis*, etc).

As stated before, irritability and gender are quite generic factors. Therefore, they may not be the best antecedents (possible causes) of diabetes from a medical point of view.

4.3. Greedy Thick Thinning network

Regarding the **parameters**, in Figure 13. GTT: conditional probability distribution of class variable, the *conditional* probability distribution of the *class* variable given the *Gender* and *Irritability* variables can be observed. Also, in Figure 14. GTT: joint probability distribution of class, Gender and Irritability variables, the *joint* probability distribution of the *class*, *Gender* and *Irritability* variables can be observed, together with the *marginal* distribution of *class* variable.

Gender	Female		Male	
Irritability	No	Yes	No	Yes
Negative	0.125	0.020833333	0.664	0.19230769
Positive	0.875	0.97916667	0.336	0.80769231

Figure 13. GTT: conditional probability distribution of class variable

Joint probability distribution:

Gender	Female		Male		Marginals
Irritability	No	Yes	No	Yes	
Negative	0.03497...	0.00186...	0.31734...	0.02939...	0.38357152
Positive	0.24479...	0.08760...	0.16058...	0.12344...	0.61642848

Figure 14. GTT: joint probability distribution of class, Gender and Irritability variables

Regarding the **structure** of the network, it can be seen in Figure 15. GTT network that the parent nodes of *class* are *Gender* and *Irritability*. It is also observed that the child nodes of *class* are *Polyuria*, *Polydipsia*, *delayed healing*, *Itching*, *sudden weight loss* and *visual blurring*. Therefore, the *class* node conditionally depends on *Gender* and *Irritability*. In turn, the *Polyuria*, *Polydipsia*, *delayed healing*, *Itching*, *sudden weight loss* and *visual blurring* nodes are conditionally dependent on *class*. Moreover, it can be seen that *weakness* is the child node of *Polydipsia*. Therefore, *class* and *weakness* are **conditionally independent** given *Polydipsia*. In addition, it has been found that the *Obesity* node is **marginally independent** of the *class* node.

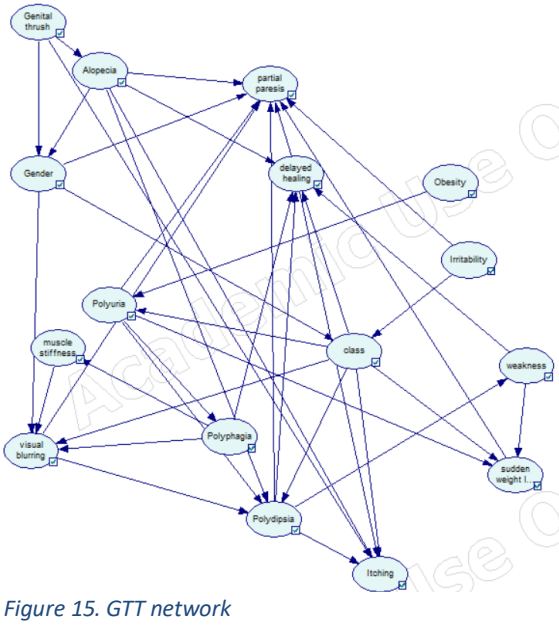

Figure 15. GTT network

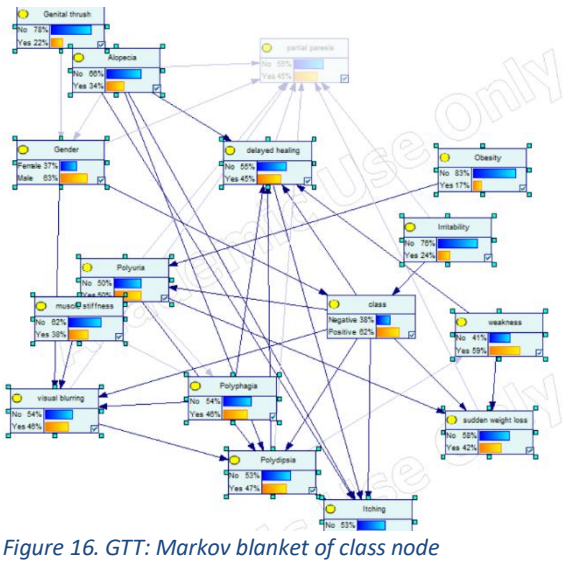

Figure 16. GTT: Markov blanket of class node

Figure 16. GTT: Markov blanket of class node shows the Markov blanket of the *class* node. It can be seen that the *class* node is **conditionally independent** from *partial paresis* given the rest of the network nodes.

As in the PC network, irritability and gender may not be the best antecedents (possible causes) of diabetes from a medical point of view.

5. Discussion

5.1. Sensitivity Analysis

Bayesian Search network

In Figure 17. BSN: Sensitivity analysis, it can be seen that the variables *Alopecia*, *Polyuria* and *Polydipsia* (also *Polyphagia*, *delayed healing* and *muscle stiffness* to a lesser extent) are **important** in this network for the calculation of the **posterior probability** distributions of *class* (target node) since that nodes are in red shades.

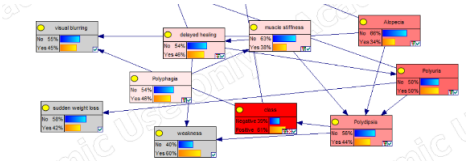

Figure 17. BSN: Sensitivity analysis

PC network

In Figure 18. PC: Sensitivity analysis, it can be seen that the variables *Irritability* and *Gender* (also *Genital thrush* to a lesser extent) are **important** in this network for the calculation of the **posterior probability** distributions of *class* (target node) since that nodes are in red shades.

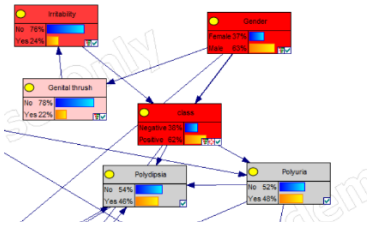

Figure 18. PC: Sensitivity analysis

Greedy Thick Thinning network

In Figure 19. GTT: Sensitivity analysis, it can be seen that the variables *Irritability* and *Gender* (also *Alopecia* and *Genital thrush* to a lesser extent) are **important** in this network for the calculation of the **posterior probability** distributions of *class* (target node) since that nodes are in red shades.

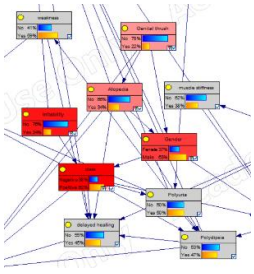

Figure 19. GTT: Sensitivity analysis

5.2.Inferences

Below are some of the inferences made for each network. Only **exact inferences** are shown since the results obtained with approximate inferences are practically identical.

Bayesian Search network

Causal inferences

$P(\text{Polyuria} = \text{Yes}|\text{class} = \text{Positive}) = 0.63$
$P(\text{Polydipsia} = \text{Yes}|\text{class} = \text{Positive}) = 0.69$
$P(\text{weakness} = \text{Yes}|\text{class} = \text{Positive}) = 0.68$
$P(\text{Polydipsia} = \text{No}|\text{class} = \text{Negative}) = 0.96$
$P(\text{Polyuria} = \text{No}|\text{class} = \text{Negative}) = 0.72$
$P(\text{visual blurring} = \text{No}|\text{class} = \text{Negative}) = 0.69$

Diagnostic inferences

$P(\text{class} = \text{Positive}|\text{Polydipsia} = \text{Yes}) = 0.97$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}) = 0.78$
$P(\text{class} = \text{Positive}|\text{sudden weight loss} = \text{Yes}, \text{visual blurring} = \text{Yes}) = 0.79$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}, \text{visual blurring} = \text{Yes}, \text{Polydipsia} = \text{Yes}) = 0.98$
$P(\text{class} = \text{Negative}|\text{Polydipsia} = \text{No}) = 0.67$
$P(\text{class} = \text{Negative}|\text{visual blurring} = \text{No}, \text{Polydipsia} = \text{No}) = 0.75$

Inter-symptom inferences

$P(\text{Polydipsia} = \text{Yes}|\text{Polyuria} = \text{Yes}) = 0.7$
$P(\text{Polyuria} = \text{Yes}|\text{Polydipsia} = \text{Yes}) = 0.79$
$P(\text{sudden weight loss} = \text{No}|\text{Polyuria} = \text{No}, \text{Polydipsia} = \text{No}) = 0.8$

From the inferences made (not all are shown in this document due to lack of space), it can be seen that **polydipsia** and **polyuria** are the symptoms most related to diabetes and it seems that they usually appear together, giving rise to the appearance of other symptoms such as sudden weight loss, visual blurring, weakness and partial paresis.

PC network

Causal inferences

$P(\text{Polyuria} = \text{Yes}|\text{class} = \text{Positive}) = 0.72$
$P(\text{Polydipsia} = \text{Yes}|\text{class} = \text{Positive}) = 0.66$
$P(\text{Polyphagia} = \text{Yes}|\text{class} = \text{Positive}) = 0.74$
$P(\text{Polydipsia} = \text{No}|\text{class} = \text{Negative}) = 0.87$
$P(\text{Polyuria} = \text{No}|\text{class} = \text{Negative}) = 0.9$
$P(\text{partial paresis} = \text{No}|\text{class} = \text{Negative}) = 0.74$

Diagnostic inferences

$P(\text{class} = \text{Positive}|\text{Polydipsia} = \text{Yes}) = 0.89$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}) = 0.92$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}, \text{Polydipsia} = \text{Yes}) = 0.97$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}, \text{Irritability} = \text{Yes}, \text{Polydipsia} = \text{Yes}) = 0.99$
$P(\text{class} = \text{Negative}|\text{Polydipsia} = \text{No}) = 0.61$
$P(\text{class} = \text{Negative}|\text{Polyuria} = \text{No}, \text{Polydipsia} = \text{No}) = 0.75$

Inter-symptom inferences

$P(\text{Polydipsia} = \text{Yes}|\text{Polyuria} = \text{Yes}) = 0.72$
$P(\text{Polyuria} = \text{Yes}|\text{Polydipsia} = \text{Yes}) = 0.76$
$P(\text{partial paresis} = \text{No}|\text{Polyuria} = \text{No}, \text{Polydipsia} = \text{No}) = 0.78$

From the inferences made (not all are shown in this document due to lack of space), it can be seen that **polydipsia** and **polyuria** are the symptoms most related to diabetes. Also, gender and irritability have a significant influence on diabetes in this model. Moreover, polyphagia seems to be a common symptom when patients have polyuria and polydipsia, whereas weakness and visual blurring are no longer so.

Greedy Thick Thinning network

Causal inferences

$P(\text{Polyuria} = \text{Yes}|\text{class} = \text{Positive}) = 0.76$
$P(\text{Polydipsia} = \text{Yes}|\text{class} = \text{Positive}) = 0.69$
$P(\text{Polyphagia} = \text{No}|\text{class} = \text{Negative}) = 0.7$
$P(\text{Polydipsia} = \text{No}|\text{class} = \text{Negative}) = 0.9$
$P(\text{Polyuria} = \text{No}|\text{class} = \text{Negative}) = 0.92$
$P(\text{Irritability} = \text{No}|\text{class} = \text{Negative}) = 0.92$

Diagnostic inferences

$P(\text{class} = \text{Positive}|\text{Polydipsia} = \text{Yes}) = 0.92$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}) = 0.94$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}, \text{Polydipsia} = \text{Yes}) = 0.97$
$P(\text{class} = \text{Positive}|\text{Polyuria} = \text{Yes}, \text{Irritability} = \text{Yes}) = 0.98$
$P(\text{class} = \text{Negative}|\text{Polydipsia} = \text{No}) = 0.64$
$P(\text{class} = \text{Negative}|\text{Polyuria} = \text{No}, \text{Polydipsia} = \text{No}, \text{Irritability} = \text{No}) = 0.84$

Inter-symptom inferences

$P(\text{Polydipsia} = \text{Yes}|\text{Polyuria} = \text{Yes}) = 0.75$
$P(\text{Polyuria} = \text{Yes}|\text{Polydipsia} = \text{Yes}) = 0.8$
$P(\text{sudden weight loss} = \text{No}|\text{Polyuria} = \text{No}, \text{Polydipsia} = \text{No}) = 0.83$

From the inferences made (not all are shown in this document due to lack of space), it can be seen that this network is very similar to the previous one. Polydipsia and polyuria are the symptoms that are most related to diabetes. Also, gender and irritability again have a notable influence on diabetes. Moreover, when patients present polyuria and polydipsia, other symptoms also appear such as weakness, visual blurring, sudden weight loss, polyphagia and partial paresis.

6. Conclusion

In conclusion, after having trained three different Bayesian networks, it seems that **gender** and **irritability** are influential on diabetes in most networks, possibly due to the fact that the dataset is not balanced (especially in the case of these two variables). Also, it seems that **polyuria** and **polydipsia** are the symptoms that most explain, or influence, or are most related to diabetes. Moreover, when these two symptoms are present, others such as weakness, visual blurring, sudden weight loss, polyphagia and partial paresis usually appear. In addition, it has been observed that the **approximate** inferences present practically the same results as the **exact** ones.

Finally, in future studies it would be interesting to balance the dataset and learn the **structure** of networks with the **aid of an expert in the domain** to obtain more realistic and efficient networks.

7. References

[1] Kumari, S., & Singh, A. (2013, January). A data mining approach for the diagnosis of diabetes mellitus. In *2013 7th International Conference on Intelligent Systems and Control (ISCO)* (pp. 373-375). IEEE.
[2] Shivakumar, B. L., & Alby, S. (2014, March). A survey on data-mining technologies for prediction and diagnosis of diabetes. In *2014 International Conference on Intelligent Computing Applications* (pp. 167-173). IEEE.
[3] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.
[4] BAYESFUSION, L. L. C. GeNIe Modeler. *User Manual. Available online: https://support.bayesfusion.com/docs/ (accessed on 21 October 2019)*, 2017.
[5] UCI Machine Learning Repository *Early stage diabetes risk prediction dataset.* https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
[6] Diabetes, World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/diabetes
[7] The Cost of Diabetes, American Diabetes Association (ADA). https://www.diabetes.org/resources/statistics/cost-diabetes
[8] Importance of Early Diabetes Diagnosis and Screening. https://www.apollodiagnostics.in/blog/importance-of-early-diabetes-diagnosis-and-screening