Practical Application 4
Machine Learning Assignment

David Cabornero Pascual
david.cabornero@alumnos.upm.es

June 18, 2022

1 Introduction

In this assignment, a dataset from the Lund University is employed in order to predict and analyze the fertility of a sample of masculine population. In fact, there are two main classes: normality and abnormality of each seminal sample.

In this case, two main Bayesian Networks will be implemented: one with exact inference and another one with approximate inference. With these tools, the dataset will be analyzed, specially the fertility variable as target variable.

2 Problem Description

Male infertility chances are higher than in females (a 16% versus a 12%) [2]. Additionally, male infertility is caused by semen quality in the 90% of the situations. Data provided for this assignment deals with the diagnosis of infertility according to the semen quality in young men.

This dataset [3] has exactly 100 samples and 9 explanatory variables. As we can see, the number of samples is specially small and we have to take it into account. Moreover, all the variables are discrete or will be discrete with some discretization operation that will be indicated properly. The description of the variables is:

- *Season*: season when the analysis is performed.
- *Age*: the age of the patient. The dataset indicates that our sample includes individuals from 18 to 36 years old, but the reality is that all the individuals are older than 27 years old. With this information, 3 groups have been created in order to discretize: 27-29, 30-32 and 33 or more.
- *ChildDisease*: binary variable that indicates a presence of a relevant disease during the childhood, such as polio, mumps or measles.
- *AccidentTrauma*: binary variable that indicates whether the individual has experimented an accident or a serious trauma.
- *Surgery*: binary variable that indicates whether the patient has undergone a relevant surgery.
- *Fevers*: it shows whether the patient has experimented fevers (1) on the last 3 months, (2) on the last year, (3) or have not experimented fevers on the last year.
- *Alcohol*: it shows the alcohol consumption habits of the individual. We have 5 possible values: hardly ever or never, once a week, several times a week, every day, several times a day.

- *Smoking*: frequency of smoking habit. It has been divided in 3 values: never, ocassionaly and daily.
- *HoursSitting*: shows the number of hours that the individual is sit per day. This variables needs to be discretized, and 4 groups have been created: 0-4 hours, 5-9 hours, 10-14 hours, more than 15 hours.

The explanatory variable is *Diagnosis*, whose values are *Normal* and *Altered*, that indicates whether the seminal sample could cause problems of fertility to the individual. In this dataset, there are 80 normal samples and 20 altered ones.

3 Methodology

For this project, two Bayesian Networks will be used in order to explain the properties of the data. The tool that will perform these tasks is *GeNIe* from *BayesFusion*.

With this kind of networks, we do not need to make predictions and calculate metrics such as accuracy or recall; we could go further. The main scope in this assignment is to find properties that explain properly the probabilistic dependencies of the diagnosis with the rest of variables. Moreover, another main objective we could have with a dataset about infertility is to determine the risk factors and its importance.

Initially, there will be certain **previous knowledge** that will be determined before learning the structure: smoking and drinking alcohol are two possible causes of infertility [4].

Furthermore, once the rest of the structure has been learned with the **PC algorithm** the data engineer must decide the direction of the edges of the *CPDAG*. In this case, the direction of the edges will be determined by causality. There are some cases that are obvious: *Age* and *Season* are always a cause and not a consequence of another variable. Consequently, these variables will be the source of the edge and not the destination. However, generally it is not as easy, but some kind of logic can be applied in general. After that, the parameters will be learned with the *Clustering* method (exact inference), which is the most accurate one. Finally, the *Likelihood sampling* method (approximate inference) will be applied in order to show the differences with the first one.

Firstly, the first relevant results will be extracted from **conditional independencies** of the *Diagnosis* variable. For this purpose, the Markov blanket is calculated and the Markov condition is applied. Moreover, the *u-separation* property is another tool that could be used to calculate more conditional independencies.

Another interesting point of view consists on calculating the best value of some variables that explains other certain fixed variables (***Maximum a Posteriori***). *BayesFusion* provides the tool *Annealed Map* in order to perform this calculations, and the posterior probability of these cases could be also guessed with this algorithm.

Finally, some simple queries will be performed to obtain certain posterior probabilities once certain values have been assumed. This will allow us to calculate the diagnosis given the causes. Because of reasons that will be justified in the next section, the distribution of the diagnosis given the season and the alcohol consumption habits will be also calculated as a table.

4 Results

The Bayesian Network that has been obtained by the PC algorithm with exact inference is in Figure 1, and the one obtained with approximate inference is in Figure 5. Furthermore, the posterior probabilities once known the *Diagnosis* variable is indicated in Figure 2. The Maximum a Posteriori estimation of the *Age*, *Smoking*, *Alcohol* and *HoursSitting* variables given the diagnosis is in Table 1. The probability of being fertile given *Alcohol*, *Smoking*, *Season* or *HoursSitting* attributes are respectively in Table 2, Table 3, Table 4 and Table 5. Moreover, some queries have been performed in Figure 3. Finally, some auxiliary graphs about the correlation of attributes are indicated in Figure 4.

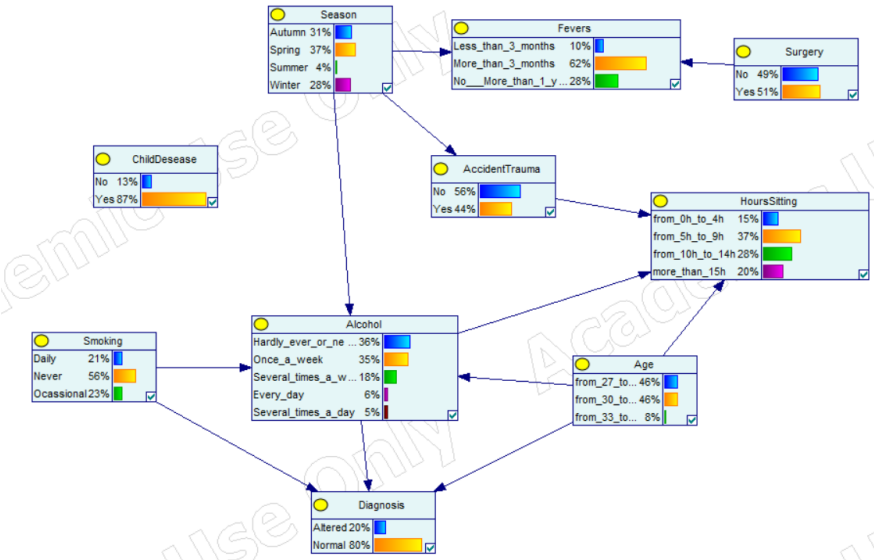
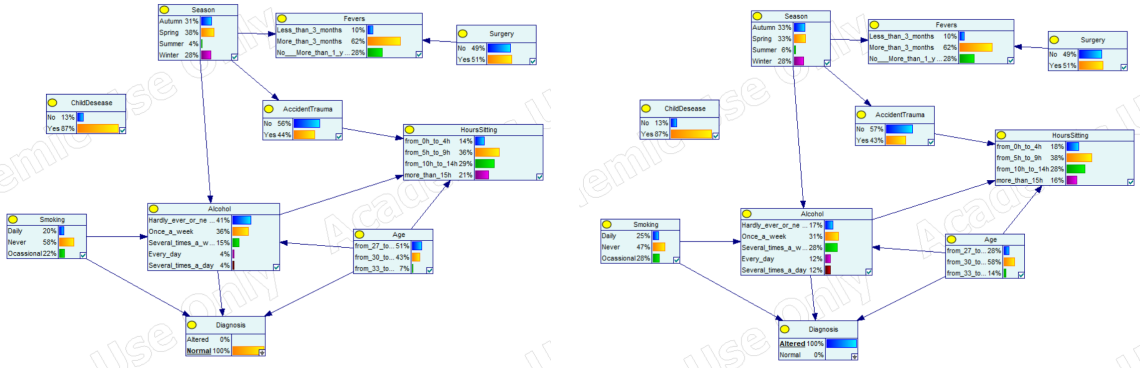


Figure 1: Bayesian network associated to the model.



(a) Diagnosis is normal.

(b) Diagnosis is altered.

Figure 2: Posterior probabilities in the Bayesian network.

Diagnosis	Age	Alcohol	Smoking	HoursSitting
Normal	26-30	Hardly ever or never	Never	15+ hours
Altered	30-33	Once a week	Never	5-9 hours

Table 1: MAP estimation of some variables given the diagnosis.

Smoking	Never	Occasionaly	Daily
Probability	83.0	75.8	76.0

Table 2: Probability (in percentage) of being fertile given the smoking frequency.

Alcohol	Never	Once a week	Several times a week	Daily	Several times a day
Probability	90.4	82.2	68.5	57.7	55.5

Table 3: Probability (in percentage) of being fertile given the alcohol consumption frequency.

Season	Winter	Spring	Summer	Autumn
Probability	79.8	81.8	70.9	78.8

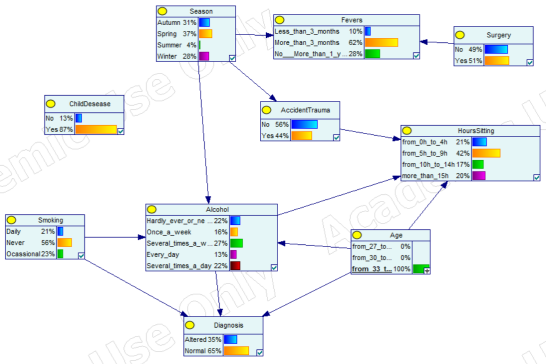
Table 4: Probability (in percentage) of being fertile given the seasonality.

HoursSitting	0-4	5-9	10-14	15+
Probability	80.2	79.2	80.2	84.1

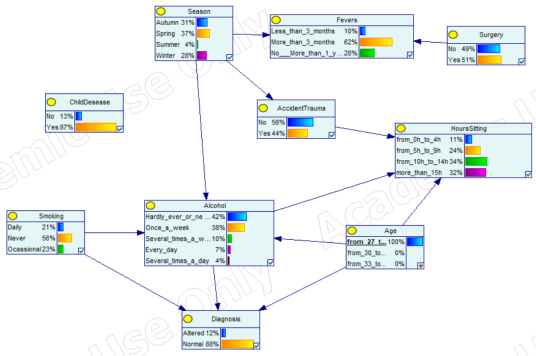
Table 5: Probability (in percentage) of being fertile given the hours that the individual is sit per day.

Season \Alcohol	Never	Once a week	Several times a week	Every day	Several times a day
Summer	84.4	77.2	72.7	53.2	51.5
All seasons	90.4	82.2	68.5	57.7	55.5

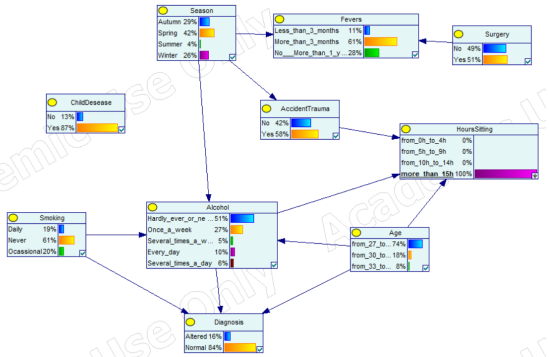
Table 6: Probability (in percentage) of being fertile given the alcohol consumption habits and whether the test was performed on summer.



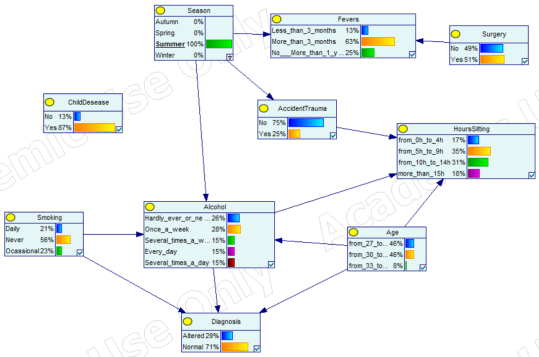
(a) Network with only older individuals.



(b) Network with only younger individuals.



(c) Network with more than 15 hours sitting.



(d) Network with only summer samples.

Figure 3: Some queries of interest.

5 Discussion

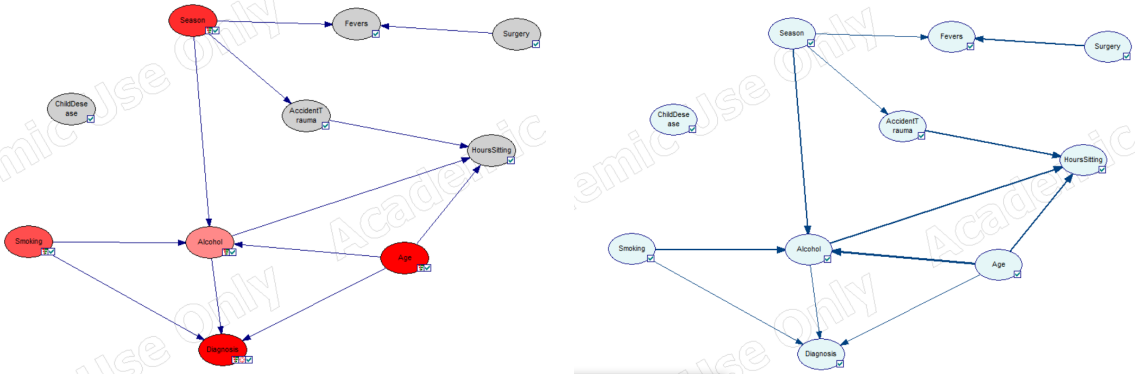
First of all, the *ChildDisease* variable must be briefly commented. The isolation of this variable implies the complete independence with any other variable of the model, so specifically we can assume that the diseases that are taken into account are not a cause of infertility.

5.1 Conditional independencies

Now, we are going to focus on the **conditional independencies**. Firstly, the Markov blanket of the *Diagnosis* variable only consists of its parents, since there is no children. Consequently, the Markov condition deductions are equivalent to the Markov blanket. They imply that, for any variable X_i (apart from *Diagnosis* and its parents), $I_P(\text{Diag}, X_i|\{\text{Smoking}, \text{Age}, \text{Alcohol}\})$. It means that, according to the model, we only need three variables to determine which is the probability of infertility: the frequency of alcohol consumption, smoking habits and age.

Secondly, we can use the u-separation property in order to find more conditional independencies with the *Diagnosis* variable. To perform this algorithm, let's see which are the nodes that must be connected with an extra edge (moralized):

- If we have *Alcohol* in the ancestral graph, *Smoking*, *Season* and *Age* must be connected by



(a) Strength of influence of each edge.

(b) Network with only younger individuals.

Figure 4: Auxiliary graphs.

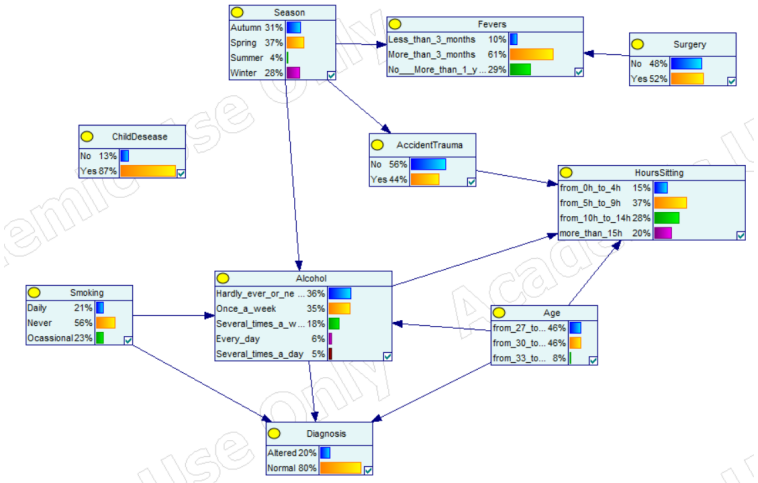


Figure 5: Same Bayesian network with approximate inference.

pairs.

- If we have *HoursSitting* in the ancestral graph, *AccidentTrauma*, *Age* and *Alcohol* must be connected by pairs.
- If *Diagnosis* is in the ancestral graph (always in our case), *Smoking* and *Age* must be connected.

Once we have this data, the most useful u-separations (which imply conditional independences) are:

- Diag $\perp$ AccidentTrauma|Season
- Diag $\perp$ Fevers|Season
- Diag $\perp$ HoursSitting|{AccidentTrauma, Alcohol, Age}
- Diag $\perp$ Surgery|{Any possible set, including the empty one.}[1]

[1]If we do not include *Fevers* on the set, the ancestral graph has two different components, one with *Diagnoses* and the another one with *Surgery*. If we include *Fevers*, every path between *Diagnoses* and *Surgery* must include *Fevers*. In both cases, *Diagnosis* and *Surgery* are u-separated by the set.

The two first u-separations explain that, once we know the season, the accidents, traumas and fevers are independent of the diagnosis. The *AccidentTrauma* case is difficult to explain, but the *Fevers* one is easier: there are a strong correlation between the seasonality and the fevers, since most individuals catch the flu, colds and another feverish diseases in winter.

The third statement asserts that we can replace the smoking habits with the presence of accidents or traumas, and the number of hours that the individual is sit will remain independent. Nonetheless, the most interesting case is the last one, since it makes sure that in every case the surgery interventions are not a relevant variable to predict the infertility of an individual.

5.2 Posterior probabilities

Now, we can learn from posterior probabilities fixing the diagnosis in Figure 2. The distributions that changed on a more significant way are the *Age* and *Alcohol* ones, so we can assume that they are the most influential variables when calculating the diagnosis.

Moreover, the **MAP estimations** in Table 1 can give us a profile of the most typical individual with infertility and without it. As we can see, this profile does not change excessively: in both cases they are people that do not smoke and they sit a great part of the day. It is true that the most common individual with infertility is older and drinks a bit more often, but the reality is that generally infertile patients have a very standard profile.

Now, we also have obtained the **marginal distribution** of the diagnosis given the frequency of alcohol and tobacco consumption, the number of hours that the individual is sit or the season when the test was performed; the rest of the subsection is dedicated to analyze these results.

The most relevant factor is the **alcohol consumption** (Table 3): drinking alcohol once a week will reduce the chances of being fertile in a 8%. Nonetheless, when the individual suffers from strong alcoholism (drink alcohol every day), chances of infertility increase a 35%, converting the alcoholism in the most determinant factor of infertility.

Smoking habits are also negative (Table 2), increasing the infertility chances in a 8% (drinking every week has the same effect as smoking). However, this fact is well known and does not contribute significantly. The real relevant fact is that smoking occasionally or daily has no significant difference regarding the infertility.

The **seasonality** (Table 4) is also an interesting factor: there is not a significant difference of fertility excluding summer, when the chances of being infertile decrease a 9%. There are several studies that handle this fact [6], and it could explain why the birth rate increases in summer and decreases in spring in many countries. Although there are some theories about the reduction of semen quality in summer (for example, the biological clock or chemical exposure), the truth is that there are not a definitive convincing theory.

Finally, regarding the **hours** that an individual is **sit per day** (Table 5) we can see that there is no significant difference between the number of hours and the chances of infertility, except when the individual is sit a great number of hours (in this case, the chances of infertility decrease). Nevertheless, several studies tend to assert the opposite statement [2], enunciating that being a long time in a seat could lead to a reduction of seminal quality.

5.3 Queries

To understand the underlying problem with the number of hours and more topics of interest, some **queries** have been performed (Figure 3). Initially, regarding the age we can see that the younger group are 23% more likely to be fertile, but we can observe also a strong relationship between the age and alcohol consumption: the older group has a probability of 44% of being alcoholic, since this probability is reduced to a 11% on the younger group. Although it is well known [2] that the age is a risk factor of infertility, we cannot assert that statement with the current dataset because of the strong correlation between alcoholism and the age.

Secondly, let's focus on the individuals that remain in a seat during more than 15 hours per day. They are more likely to have a good seminal quality, but the reason underlying is again a lurking variable: this small set of individuals has a tendency to be young, drink alcohol with less frequency and smoke less.

Finally, the results obtained in summer have the same problem: there is a probability of 30% of being alcoholic, which may decrease the global chances of being fertile during summer. For this reason, Table 6 has been performed, to allow the reader to compare the general distribution of diagnoses in summer and in general taking into account the grade of alcoholism. As we can see, there is still a significant difference in summer, but it has been decreased to a 4%.

5.4 Sensitivity and influence

In addition with the previous information, *BayesFusion* provides some tools that allow the users to explore more deeply the relationships between variables. In Figure 4, we can see two graphs: one of influence and another one with a sensitivity analysis.

In the **influence** graph, we can appreciate how strong are the relationships between the variables: the higher correlation between two nodes, the wider edge in the graph. As we can see, the *HoursSitting* variable is strongly dependent of the alcohol consumption and age, which we previously predicted in the queries. The seasonality is also correlated with the alcohol consumption, which was predicted also (in summer, the chances of alcoholism increase significantly). In addition, there is also a strong correlation between the alcohol consumption and smoking, which lead to think that there is an important sector of the population that consume both drugs or neither of them. We can easily contrast this hypothesis looking at the *MAP* estimates of the *Smoking* variable:

- People who never smoke also never (or hardly ever) drink with a probability of 49%.
- People who smoke occasionally drink alcohol once a week with a probability of 48%.
- People who smoke every day drink alcohol once a week with a 34%.

Indeed, individuals who smoke are more likely to drink alcohol and there is another important group of *sobers*.

Finally, the **sensitivity** graph show the influence that each variable have in the diagnosis. According to that, the most important variables to predict infertility (in descending order) are the age, seasonality (specifically whether it is summer), smoking and alcohol consumption. In addition, although it was already mentioned, we can remark again that the number of hours that the individual is sit per day is not relevant in the model.

5.5 Approximate inference

As we can see in Figure 5, the approximate algorithm does not change specially the marginal distributions. The analogous tables have not been included, because it would have caused repetition since every value has generally less than a 1-2% of difference with the exact inference.

Why is this difference so low? Many of approximate inference methods are based on the Monte Carlo method. By default, this method samples 10.000 cases, which is clearly enough with our dataset composed of 100 cases. The number of samples could be reduced in order to present a model that differs more from the first one, but in this case we are only introducing random noise. Consequently, we will not gain extra information doing approximate inference, we will only obtain a less accurate network.

6 Conclusion

In this project, a dataset of fertility prediction on men has been handled. For this purpose, a Bayesian network has been performed with the PC algorithm (exact inference) and the likelihood sampling method (approximate inference). Before learning the structure, some initial information has been given to the algorithm: smoking and alcohol consumption cause infertility.

Mainly, the network detected as **risk factors** alcoholism, smoking and age of the individuals. The rest of the attributes are independent of the diagnosis if we know these three facts. Nonetheless, the seasonality is also relevant, and knowing this attribute is enough to cause conditional independence with some of the variables. This strange dependence between the season and the diagnosis is caused because of summer: infertility is significantly more common on this month.

The network has detected other variables that are **completely independent** of the diagnosis: knowing whether the individual has suffered from some relevant childish diseases or has undergone a surgery are not relevant facts. Additionally, recent fevers, accidents, severe traumas, or a great number of hour per day in a seat are not much relevant according to the model, although there are some studies that disagree with the last case.

Moreover, alcoholism is the worst risk factor: drinking every day decreases the chances of being fertile more than a 30%, and drinking once a week decreases them a 8%. In addition, smoking decreases this chances another 8%, and being in summer decreases them a 4% (this could explain the fact that the birth rate increases in summer and decreases in spring).

Finally, it is remarkable that the **most common** fertile individual does not smoke or drink and is young, but the most common infertile individual does not smoke either, drinks only once a week and is quite older. Thus, we could assert apparently that people who come to the doctor with a problem of infertility is usually a healthy individual.

References

[1] C. Bielza Lozoya and P. Larrañaga Mújica. *Data-driven computational neuroscience : machine learning and statistical models*. Cambridge University Press, Cambridge, 2020.

[2] R. M. Blay, A. D. Pinamang, A. E. Sagoe, E. D. A. Owusu, N. K.-K. Koney, and B. Arko-Boham. Influence of lifestyle and environmental factors on semen quality in ghanaian men. *International Journal of Reproductive Medicine*, 2020:1–7, 2020. Online, accessed 13-Jan-2022. Link: https://www.hindawi.com/journals/ijrmed/2020/6908458/.

[3] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres, and M. Johnsson. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16), 2012. Online, accessed 09-Jan-2022. Link: https://archive.ics.uci.edu/ml/datasets/Fertility.

[4] Kolesnikova, L.I., and et al. Causes and factors of male infertility. *Annals of the Russian academy of medical sciences*, 70(5):579–584, December 2015. Online, accessed 09-Jan-2022. Link: https://vestnikramn.spr-journal.ru/jour/article/view/551.

[5] K. B. Korb and A. E. Nicholson. *Bayesian artificial intelligence*. CRC Press, 2011.

[6] R. J. Levine. Seasonal variation of semen quality and fertility. *Scandinavian Journal of Work, Environment and Health*, 25:34–37, 1999. Online, accessed 13-Jan-2022. Link: https://www.jstor.org/stable/40966973.