Revue Française d'Economie et de Gestion
www.revuefreg.fr
Volume 1 : Numéro 1

The use of Bayesian Networks in the prediction of bankruptcy

L'utilisation des réseaux bayésiens dans la prédiction de la faillite

Madiha ZAMMEL
PhD Student at the Faculty of Economics and Management of Sfax
Governance, Finance and Accounting Laboratory.
madiha-zammel@hotmail.fr

Walid KHOUFI
Professor at the Institute of Higher Business Studies of Sfax.
Governance, Finance and Accounting Laboratory.
walid.koufi@yahoo.fr

Date de soumission : 29/04/2020
Date d'acceptation : 12/06/2020

Pour citer cet article :
ZAMMEL M. & KHOUFI W. (2000) «The use of Bayesian Networks in the prediction of bankruptcy», Revue Française d'Economie et de Gestion «Volume 1 : Numéro 1» pp : 60- 88

Abstract

The purpose of this paper is to compare the development of Bayesian Network models: naïve Bayesian network (NBN) and a maximum weight spanning tree (MWST) in forecasting the failure of the Tunisian companies. Using a sample of 130 small and medium Tunisian companies and a battery of 9 financial ratios calculated for the 2005-2012 period, it can be concluded that, for a classification problem, supervised learning with naïve architecture is more appropriate and gives more relevant results than an unsupervised learning model with a maximum weight spanning tree.

Keywords:

Company's failure, forecasting, Naïve Bayesian network (NBN), maximum weight spanning tree (MWST).

Résumé :

Le but de cet article est de comparer le développement des modèles de réseau bayésien: un réseau bayésien naïf (NBN) et un arbre couvrant de poids maximum (MWST) dans la prévision de la défaillance des entreprises tunisiennes. En utilisant un échantillon de 130 petites et moyennes entreprises tunisiennes et une batterie de 9 ratios financiers calculés pour la période 2005-2012, on peut conclure que, pour un problème de classification, l'apprentissage supervisé avec une architecture naïve est plus approprié et donne des résultats plus pertinents que un modèle d'apprentissage non supervisé avec un poids maximal couvrant l'arbre.

Mots-clés:

Défaillance de l'entreprise, prévisions, réseau bayésien naïf (NBN), arbre couvrant le poids maximum (MWST).

Introduction

Logically, business failure is a phenomenon that enables to purify the inefficient and unproductive business market. However, this reality is costly for all the stakeholders: the creditors are likely to remain unpaid, the employees may lose their jobs, and the homeowners may lose their funds…. For this reason, acting before the outbreak of the problem and avoiding the risk of failure certainly saves the costs, that is why forecasting plays a very important role in the treatment of failure not only at the theoretical level in studies and academic research, but also in practice.

Failure prediction, which deals with the assessment of the companies' default risk on the basis of the accounting data, has long been at the core of the researchers' interest since the beginning of the 20th century with the research of Rosendale (1908), Fitz Patrick (1932). Failure prediction consists in using statistical tools to detect the appearance of financial signals that indicate a forthcoming failure. Failure prediction models are multiple. Generally, the common aim of these models is to assign any company to one of the two groups (the group of failing firms, or the group of healthy companies) using accounting ratios.

Although the methodology used in these models and the analysis variables have not much changed, the evolution of statistical analysis tools makes this topic still relevant. Indeed, the history of evolution of predictive models can be divided into two major periods: the period of traditional models and that of the artificial intelligence models. Traditional studies of bankruptcy prediction can be classified into two groups according to the existence or absence of hypotheses about the distribution of the variables. If the distribution law of the variables is known in advance, we can talk about parametric statistical methods, otherwise, we talk about nonparametric statistical methods.

In the first instance, the parametric methods of the statistical classification are intended to establish a functional relationship the form of which is given a priori between an endogenous variable and exogenous variables in which the distribution law is supposed to be known. The parametric prediction models can be grouped into three families. Initially, the authors used a single variable to explain the business failure where we can talk about a univariate statistical analysis (Fitz Patrick (1932), Winakor and Smith (1935), Merwin (1942), Beaver (1966)).

The number of variables multiplies assuming they follow the normal distribution, then we start talking about a multivariate discriminate analysis (Altman (1968)). Finally, the multi-normality variables hypothesis is rejected in favor of the hypothesis which states that the explanatory variables have different distributions, hence, the birth of the logistic regression methods (Probit/Logit) (Ohlson (1980)). When the financial ratios do not follow a multi-normal distribution and the errors follow neither the normal distribution nor the logistic law, the business failure prediction using parametric templates becomes impossible, hence, the resort to the nonparametric models requiring no hypotheses about the distribution of the variables.

Then, the nonparametric techniques can be classified into two categories: models based on recursive partitioning technique (Frydman et al. (1985), Bardos (1989)) and models based on kernel estimator technique (Calia and Ganugi (1997)).

The second stage in history started with the artificial intelligence algorithms that were developed through many constraints related to traditional statistical methods. The Bayesian Networks method is among the most famous techniques in this line of research. Artificial intelligence techniques give significant results about business failure forecasting without requiring any statistical restriction. The many studies that focused on the comparison between these techniques and their traditional counterparts confirm their superiority in terms of prediction quality (Mamoghli and Jackson (2006)). More artificial intelligence techniques have the advantage to be adapted to any problem, among other things, the business failure prediction.

Regarding the Tunisian context where the companies' failure followed an upward curve, we proposed to apply the naive Bayesian network model (supervised learning) and the maximum weight spanning tree (Bayesian network model with unsupervised learning) to predict the companies' failure using a sample of 130 companies. For this purpose, the second section of our contribution is devoted to the presentation of the Bayesian network model. The third section is reserved for the application of this model to the prediction of failure of Tunisian companies.

1. Presentation of the Bayesian Network Model

The Bayesian networks represent a modeling tool that links the probability theory to that of the graphs. This type of network makes it possible to describe the relationship between the random variables through a graphical representation of knowledge and carry out a probabilistic reasoning on them, Hence, this network is often called the probabilistic oriented graphical models, probabilistic networks, belief networks, or probabilistic independence networks.

This type of network enables to quantitatively and qualitatively describe the functioning of a system through the representation of the relationship between the variables and the calculation of the different probabilities between them. The Bayesian networks are probabilistic graphical models that help compute the conditional problems on the basis of the Bayes' theorem. Actually, Thomas Bayes developed a theorem based on the information dissemination within the network. In other words, it is a post-calculation of the problems of some variables on the basis of a number of observations on other variables (this is the conditional probability).

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

With P (A) = the probability of achieving event A

P(B) = the probability of achieving event B

P (A/B) = is the posteriori probability of A given that B is true. It means the probability of achieving A event (which is directly dependent on B) given that B is achieved.

P (B/A) = this is the likelihood function of A for a known B

The Bayesian networks describe the distribution of the probabilities associated with a set of variables some of which are directly dependent and others are conditionally independent. In order to better model the uncertainty that links the different variables to be studied, the graph of a Bayesian network visually represents the dependence and independence relationship between the variables of the model and then enriches this graphic structure through the quantification of these relations using the calculation of the probabilities.

Two main families of Bayesian networks can be studied: the dynamic and the naive Bayesian networks. Firstly, a dynamic or temporal Bayesian network (often called DBN for Dynamic Bayesian Network) is a Bayesian network integrating the notion of time. In other words, the value of a variable can be changed from one instant (t) to another (t + 1). Several authors have been interested in the definition of this type of network. For example, Neopolitan (2004) defined a dynamic Bayesian network as an extension of the Bayesian networks, which is modeled on a temporal process. On his part, Murphy (2002) defined it as a special Bayesian network used for stochastic dynamic process models.

If we consider a set of variables that evolve over time:

$$X(t) = X_1(t), X_2(t), \ldots, X_n(t)$$

A dynamic Bayesian network represents the conditional probability distribution of these variables for a bounded interval [0; T]. A dynamic Bayesian network is represented by a network the structure of which is duplicated for each time step. In this graph, each node represents a random variable the value of which indicates the state occupied by this variable at each instant (t). Next, the naive Bayesian networks are generally used to solve classification problems. In this case, the model requires a number of possible classes "C" and the observed variables "X1, ..., X i" which characterize each of these classes: for example, within a population of firms, the individuals are sorted into two classes; Healthy and failing, depending on the observed symptoms, such as the decline of the turnover, insolvency.... The structure of a naive Bayesian network is composed of only two levels. The first is made up of a single node A whereas the second includes several nodes with only one parent A.

2. The failure prediction of the Tunisian companies through the Bayesian network model

3.1. Research methodology

3.1.1. The objective of the study

The main objective of this article is to apply the Bayesian network model (with supervised learning), which is a naive model in the prediction of the Tunisian firms' failure, then compare the performance of this model with that of the unsupervised learning model: the maximum weight spanning tree (MWST model).

3.1.2. The sample

In the context of this study, we prepared our own database, which includes 130 small and medium Tunisian enterprises 65 of which are failing and 65 healthy. During the formation of the sample, we have met certain criteria of homogeneity. Firstly, among non-financial companies, only firms belonging to the non-financial sector are retained, that is the sector homogeneity.

Afterwards, the firms that make up our sample are the small and medium-sized enterprises (SMEs), the thing which brings us back to talk about the size homogeneity. To define an SME on the basis of the data availability, we have adopted the size criterion in accordance with some earlier researchers, such as Khoufi and Feki (2008). Therefore, an SME is a company the capital of which does not exceed 1 million Tunisian dinars. Finally, the homogeneity of the groups of companies is respected by using the combined sampling technique. Indeed, this technique consists in associating with every failing firm a healthy company of the same size and sector of activity. The sub-sample of healthy firms is composed of active companies that have serious financial problems. However, to form the sub-sample of the failing firms and due to the multiplicity of definitions of the concept of failure, we define a default company is any company that initiates legal proceedings.

The financial statements (balance sheets and income statements) of these companies are collected from services from Central Bank, the INS and the control offices. The financial statements collected for each company cover 3 years (the year of the failure and the previous two years) for the 2005/2012 period. In this way, a three-year database is established involving 130 companies.

To develop our model, we established an endogenous (explained) variable and several exogenous (explanatory) variables. Our endogenous variable Y represents the company's financial situation which may be failing or healthy. It takes value 1 if the company belongs to the group of failing firms and 0 (zero) if it is part of the healthy firms.

3.1.3 . Definition and measurement of the variables

Through the review of previous research about the failure topic, (Altman (1968), Ohlson (1980), Hopwood and al. (1994), McKee and Lensberg (2002), Lili. S and Prakash. P. S. (2007), Salehi. M and Abedini. B (2009), Suzaida. B and Muhammad Z. AH (2012)...), we have established a battery of 32 financial ratios used as explanatory variables. These ratios are grouped into 4 groups: the liquidity ratios (12 ratios), the management ratios (8 ratios), the profitability ratios (3 ratios) and the structure ratios (9ratios). The initial list of 32 calculated ratios is presented in appendix 1.

Indeed, due to the large number of financial ratios that can be used as indicators of failure, an appropriate selection of variables is necessary to better develop the model. For this reason, we chose to reduce our initial list of financial ratios. The selection of the explanatory variables used in this study is made using a heuristic method based on the analysis of the partial correlation. This heuristic, previously used by Lili and Prakash (2007), is based on the assumption that the dependence between each pair of variables is linear and can be measured by the correlation coefficient.

Actually, as an initial step, this method consists in calculating the correlations between the variables (32 financial ratios and the endogenous variable (Y)). In fact, only the variables that have a significant correlation (P-value of Pearson's correlation ≤0, 01, that is, the significance at 1% level) are considered to be dependent. Firstly, we examined the significance of the correlations at the threshold of 1%, 5% and 10%. Then, we chose to retain the threshold of 1% for the rest of our work because it actually guarantees an appropriate functioning of the model using the minimum number of variables. In fact, the choice of the threshold of significance varies from one study to another. For example: Lili and Prakash (2007) used a threshold of 10% while Suzaida. B and Muhammad. Z. AH (2012) used a 5% threshold. The choice of the significance optimal threshold is a line of research in itself. The significance analysis of the correlation (at the threshold of 1%) between the exogenous variables and the endogenous one gives us that only 10 variables (R7, R10, R15, R16, R19, R20, R21, R23, R28, R30) are connected to 'Y' (the situation of the failing or healthy company). Using these variables, we were able to create 45 pairs of variables.

The second step of our heuristic method consists in examining the partial correlation. Indeed, to avoid the distribution of information in the calculation and use of less relevant variables, we proceed to analyze, in each pair, if a variable is dependent on 'Y' in relation to the other variable of the pair by examining the partial correlations between this variable and 'Y' and by controlling the other variable of the pair. For example, the analysis of the pair formed by R7 and R16 gives the following result. The significant partial correlation between the Y and R7 variables (0.006), given R16, means that the R7 variable has an additional contribution to the prediction of the Y variable given R16, the significant partial correlation between Y and R16 (0.000) given R7, means that R16 has an additional contribution to the prediction of Y given R16. As a result, the R7 and R16 variables are held as explanatory variables. By applying the same reasoning with the 44 remaining pairs (each variable has an additional contribution to the prediction of Y given the other variable of the pair). It is noted that some partial correlations are significant (at the threshold of 1%), while others are not. Given that the R23 ratio has no significant correlation, therefore, they are eliminated from the list. Finally, the significant partial correlations (15 pairs) are presented in Appendix 2.

Finally, the ratios which are used as explanatory variables are: R7 (the general financial autonomy ratio); R10 (the indebtedness ratio); R15 (the rate of equipment renewal); R16 (the coverage rate of personnel costs); R19 (the ratio of the importance of the assets depreciation expenses); R20 (the importance ratio of the financial expenses); R21 (financial profitability ratio); R28 (the long and medium-term indebtedness ratio) and R30 (the financial balance ratio). The descriptive analysis of these variables is presented in appendix 3.

3.2. The model presentation

The model that we intend to develop, in the context of this article, is the model of the Bayesian network. It is a probabilistic graphical model that represents a set of the random variables on a given problem and the different relationships between them. Since the problem is about the classification of businesses according to their situation failed or healthy, we chose to apply a naïve Bayesian network model with exact inference method, which is the most used in this kind of problem.

In a naïve Bayesian model, the node that represents the problem is the root node (i.e. node without a parent). In the context of a business failure, the root node represents the company's situation (the failing or healthy) and the other nodes represent the selected ratios (the explanatory variables). A Bayesian network still works with discrete variables and, if ever there are continuous variables, they must be transformed, this is what is called the discretization phase.

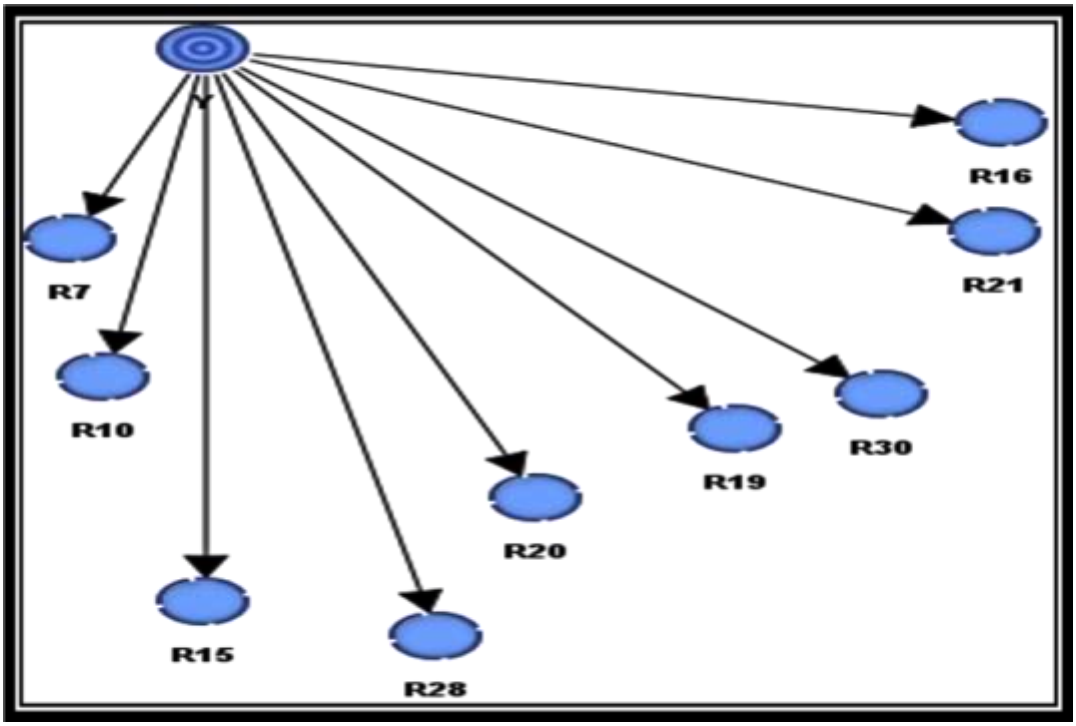
Several methods are available for the discretization. In our study, we chose to use the method of equal distances with interval of 3 because it is the most used method in classification problems where the variables are continuous. This method consists in dividing the equal values of each variable into three intervals. The table in appendix 4 shows the results of discretization.

3.3. Analysis and interpretation of the results

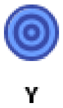
3.3.1. The construction of the model

The BayesiaLab software helps with the learning of the Bayesian network by taking the whole of the discretized database as the input model without data sampling. The structure of the naïve Bayesian network with 9 selected ratios is presented in the following figure:

Figure 1: Graph associated with the naïve Bayesian network model



In this graph, there is a set of nodes and arcs with:



: This is the target node (parent node) that represents the situation of the failing or healthy company)

Represent the child nodes (leaves), which are the financial ratios.

The arcs that represent the correlations between the variables

This graph shows that the company's situation can be viewed on the basis of its general long and medium-term indebtedness level, its general financial independence, its equipment renewal rate, the importance rate of its personnel costs, the degree of importance of its asset depreciation expenses, the degree of importance of its financial expenses and the level of its financial equilibrium.

3.3.2. The result interpretation

The BayesiaLab software offers different methods of the result analysis on the basis of the obtained network.

3.3.2.1. Analysis of the discovered relationships

The relationships in a naïve Bayesian model are oriented from the root node towards the child nodes (leaves). The BayesiaLab offers 3 different measures for each relationship: the distance of Kullback-Leiber, the relative weight and Pearson's correlation.

First, the Kullback-Leiber distance and the relative weight are two measures indicating the strength of the relationship and the level of correlation between the variables. In fact, the more Kullback distance -Leiber and the relative weight converge towards 1, the stronger the relationship between the variables and vice versa. Then, Pearson's correlation refers to the direction of the correlation between the variables and its degree of significance.

The analysis of the relationships between the variables of the built naïve Bayesian network is presented in the following table:

Table 1: Analysis of the relationship of the naïve Bayesian network

Parent	Child	Kullback-Leibler divergence	the relative weight	Pearson's correlation	p-value
Y	R15	0,687	1	-0,256	0,000***
Y	R10	0,669	0,790	0,178	0,000***
Y	R7	0,663	0,787	-0,178	0,000***
Y	R21	0,497	0,663	-0,255	0,060*
Y	R28	0,387	0,593	0,155	0,000***
Y	R16	0,285	0,455	0,177	0,042**
Y	R30	0,262	0,418	-0,157	0,008***
Y	R19	0,104	0,165	-0,096	0,059*
Y	R20	0,032	0,051	0,057	0,422

***: significant at 1% threshold, **: significant at 5% threshold and *: significant at 10% threshold

Table 1 analyses the level and the nature of the relationship between the company's situation and its indicators (ratios). The results show the presence of a strong (Kullback-Leiber=0,627 and relative weight=1) and negative relationship (= α = - 0,256 Pearson's correlation) between R15: the equipment renewal rate and the company's situation (Y), which means that the latter is heavily affected by the degree of the equipment aging. In other words, the non-renewal of equipment (weak R15) makes the business unable to follow technical progress and acquire new technologies to deal with the increasing competition, which can lead to failure. In fact, the analysis of R19 ratio (the importance of the asset depreciation expense) confirmed this result. On the other hand, the relationship between Y and R19 is weak (= 0.104 Kullback-Leiber) and the relative weight (= 0,165) is negative (α = - 0,096). This relationship can be interpreted as follows: If R19 is high, this means that the level of company's asset depreciation is high, which helps us declare that the company's materials and equipment are frequently renewed and therefore, the company attaches a great importance to technical progress which makes it always about to acquire sophisticated equipment.

The relationship analysis test showed that financial profitability (R21) has a moderately negative correlation with the company's situation (Kullback-Leiber = 0, 497, relative weight = 0,663 and α = - 0,255). This means that higher the company's financial profitability is, of the lower its failure probability will be.

The analysis of the relationship between Y and R10 (indebtedness ratio) as R28 (medium and long-term indebtedness ratio) showed a strong correlation for R10 (Kullback-Leiber=0,669 and relative weight=0,790) and a moderately strong one for R28 (Kullback-Leiber=0, 387 and a relative weight= 0,593). This relationship is positive for both variables (αR10 = 0,178 and αR28 = 0.155), which means that, the more the company's financial structure is based on indebtedness (high indebtedness level), the higher its failure probability will be.

In the same direction, a high indebtedness level means that the company is closely linked to its bank, hence the weakness of its financial autonomy and the increase of its failure probability. This idea is supported by the analysis of the relationship between Y and R7 (financial autonomy ratio), which is a strong (Kullback-Leiber = 0, 663 and relative weight = 0,787) and negative (α = - 0,178) relationship. The importance rate of the personnel costs (R16) moderately affect (Kullback-Leiber = 0, 258 and relative weight=0,455) and positively (α=0,177) the company's situation. Indeed, the more important the personnel costs are, the higher the failure probability will be.

Let us apply the same reasoning to the analysis of the relationship between R20 (ratio of the importance of the financial expenses and Y as a positive relationship (α=0.057), which makes us believe that the company's failure probability increases with the rise of the financial charges.). Finally, the financial equilibrium (R30) negatively affect (α= -0.157) the company's situation. In fact, the higher this ratio is (the company is in a financial equilibrium), the lower the failure probability will be.

3.3.2.2. the target analysis: the company's situation

The analysis of the target variable consists in examining its relationship with the other variables according to each possible modality.

The target analysis report under request provided by BayesiaLab is used to measure this relationship by means of the binary mutual information, the binary relative importance and the modal value. Firstly, the mutual information of binary helps measure the information contribution of a variable to the target value. Then, the relative binary importance measures the importance of the variable in relation to the target (Y) variable. Finally, the modal value is the average value of the independent variable for each value of the target variable.

It should be recalled that our explanatory variables are continuous, and that the method of equal distances with 3 intervals is used for discretization, hence, three modalities are obtained for each variable on the basis the chosen intervals (low, medium, high) therefore, the discretization of R7 gives us:

Weak (1/3)	Medium (2/3)	High (3/3)

-6,82 -4,24 -1,66 0,92

These different modalities are going to be used in the analysis of the modal value of each explanatory variable. The analysis report of the target variable gives us the results presented in table 2.

Table 2: Importance of the nodes in terms of the available information about the companys situation

Y= healthy (50.3817%)				
Node	**Binary mutual information**	**Binary relative importance**	**Modal value**	
R15	0,0627	1	(2/3) moyen	46,97%
R10	0,0593	0,8901	(1/3) faible	84,34%
R7	0,0571	0,8527	(3/3) élevé	84,12%
R21	0,0507	0,7933	(3/3) élevé	92,65%
R28	0,0497	0,6921	(1/3) faible	89,87%
R16	0,0285	0,4549	(1/3) faible	91,41%
R30	0,0262	0,4183	(3/3) élevé	95,45%
R19	0,0104	0,1654	(3/3) élevé	76,93%
R20	0,0032	0,0506	(1/3) faible	98,99%

Y = failing (49,6183)%				
Node	**Binary mutual information**	**Binary relative importance**	**Modal value**	
R15	0,0627	1	(2/3) moyen	38,46%
R10	0,0593	0,8901	(3/3) élevé	92,82%
R7	0,0571	0,8527	(1/3) faible	90,76%
R21	0,0507	0,7933	(1/3) faible	63,59%
R28	0,0497	0,6921	(3/3) élevé	63,59%
R16	0,0285	0,4549	(3/3) élevé	80,51%
R30	0,0262	0,4183	(1/3) faible	94,87%
R19	0,0104	0,1654	(2/3) moyen	97,95%
R20	0,0032	0,0506	(3/3) élevé	98,97%

This table shows that the most important nodes in the explanation of the company's situation are the equipment renewal rate (binary relative importance=1), the indebtedness ratio (binary relative imprtance=0, 8901) the general financial autonomy ratio (binary relative imprtance=0, 8527). Then, the other explanatory variables are; the financial profitability ratio (binary relative importance=0, 7933), the long and medium-term indebtedness ratio (binary relative importance=0, 6921), the importance of personnel ratio (binary relative importance=0, 4549), the financial balance ratio (binary relative importance=0, 4183), the importance of the depreciation expense ratio (binary relative importance=0, 1654) and the importance of the financial cost ratio (binary relative importance = 0, 0506).

Moreover, the results show that an average level of equipment renewal with a 46,97% probability added to 84,34% of a low indebtedness level, 84,12% of a high level of financial autonomy, 92,65% of a high level of cost-effectiveness, 89,87% a low level of long and medium-term indebtedness, 91.41% a low level of personnel costs, 95,45% a high level of financial equilibrium, 76,93% of a high level of depreciation expenses 98,99%, a low level of financial charges, improve the company's situation and protect it against failure with a probability of 50,3817%.

On the other hand, failure may occur with the probability of 49,6183% if the explanatory variables come together in the following way: 38.46% of an average level of renewal of equipment, 92,82% of a high level of indebtedness, 90,76% of a low level of financial autonomy, 63.59% of a low level of financial profitability. 63.59% of a medium and long-term high level, 80, 51% of a high importance of personnel costs, 94,87% of a low financial balance, 97.95% of an average level of depreciation expenses, and 98,97% of a high level of finance charges.

3.3.2.3. Analysis of the model overall performance

After examining the different relationships found in the model as well as the impact of the explanatory variables on the target variable (the company's situation: failing or healthy), it is appropriate to analyze the model classification capacity as well as its overall performance.

In fact, the classification errors are of two types: there are companies that are actually flawed but are classified as healthy (type 1 error) and there are healthy companies but are classified in the group of failing firms (type 2 error). The 'analysis of targeted performance' function enables us to consult the BayesiaLab software about the classification errors of our model.

Table 3: The NBN confusion matrix

	Healthy (Y=0) 195	Failing (Y=1) 195
Healthy (Y=0)	181 (92,82%)	20 (10,26%)
Failing (Y=1)	14 (7,18%)	175 (89,74%)

Table 3 shows that our model can correctly classify 89,74% of the failing firms (standard error 1 = 10,26%) and 92,82% of the healthy ones (standard error 2 = 7.18%). Overall, the good ranking rate of this model (the relationship between the number of correctly classified companies and the total number of companies) is estimated at 91,28%. The ROC index analysis makes us draw a conclusion about the model performance: the ROC index amounts to 84,96%.

3. Naive Bayesian network versus the maximum weight spanning tree

In this section, we will change the learning algorithm and check if the predictive ability of the model will be changed later. Insofar as the naïve Bayesian network is a supervised learning model, we'll choose at this stage to test a model of unsupervised learning: the maximum weight spanning tree (MWST).

4.1. Presentation of the MWST model

This model is inspired by the studies of Chow and Liu (1968) who proposed a research-based method about the maximum weight spanning tree (MWST). This method, which consists in reducing the graph space to the tree space, can be directly applied to the search for the structure of a Bayesian network.

The first thing to do is to set a weight to each potential edge R_i–R_j of the tree. This weight can be the mutual information between the R_i and R_j variables in accordance with the work of Chow and Liu (1968), or even the variation of a local score when R_i is chosen as the parent of R_j as suggested by Heckerman, Geiger and Chickering (1994).

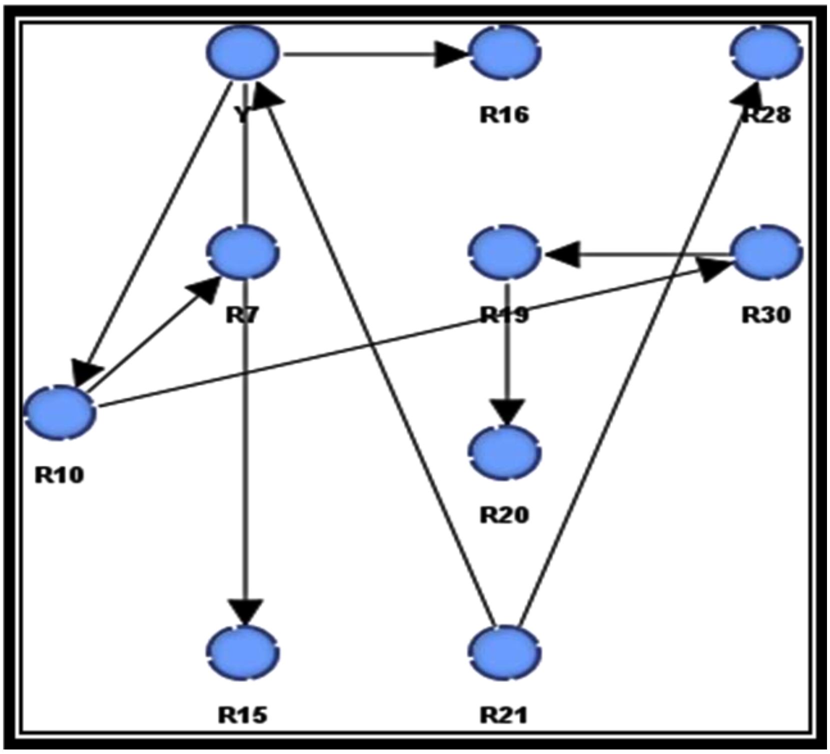
Then, once this weight matrix is defined, it would be necessary to use one of the standard algorithms to solve the problem of the maximum weight spanning tree, such as the Kruskal algorithm or that of Prim to fix the model of the undirected tree. Finally, this created undirected tree should be directed by means of a root node, then by browsing and directing the tree through in-depth research (from the outside towards the inside). The root node can be chosen either randomly, by using a priori knowledge, or by taking the variable representing the class for classification problems.

4.2. The implementation of the MWST model

To implement the MWST model, we applied the same database and both the explained and the explanatory variables that we used for the training of the naive Bayesian network model, as well as the same descritisation method. The used inference is an exact inference. Firstly, we tested two models by always changing the learning algorithm (Kruskal or Prim) and found that the model using the Kruskal algorithm gives the best results. Actually, the Kruskal algorithm builds a maximum weight spanning tree by first evaluating all the weights of the graph then, by sorting the edges per descending weight if they do not create a cycle.

The maximum weight spanning tree obtained by setting Pearson's correlation as a learning score is presented in figure 2.

Figure 2: Graph associated with the MWST model



The analysis of the targeted network performance gives us the following confusion matrix:

Table 4: Confusion matrix for the MWST model

	Healthy (Y=0) 195	Failing (Y=1) 195
Healthy (Y=0)	171 (87,69%)	90 (46,16%)
Failing (Y=1)	24 (12,31%)	105 (53,84%)

According to this matrix, the overall good ranking rate of this model accounts for 70,76%. Moreover, the overall ROC index for this model is at the threshold of 72,65%

4.3. Comparison between an NBN and an MWST

When comparing the results of both NBN and MWST studied models, we can observe, as a first difference, that a priori knowledge of the variable used for classification does not apply in the MWST approach, unlike the structure presented by the naïve Bayesian network model that requires the prior setting of a target for classification.

The results of the comparison between both models are resented in the following table

Table 5: Performance of both models

	Type 1 error	Type 2 error	Good ranking rate of failing companies	Good ranking rate for healthy companies	Overall Good ranking rate	ROC index
NBN	10,26%	7,18%	92,82%	89,74%	91,28%	84,96%
MWST	46.16%	12,31%	53.84%	87,69%	70.76%	72,65%

This table enables to focus on the high classification capacity of the failing companies for a naïve Bayesian network model which amounts to 92,82 per cent against 53,84 per cent for the maximum weight spanning tree. Moreover, the overall good ranking of the naïve model rate (91,28%) is higher than that of MWST (70.76%). Therefore, these results make us conclude that for a classification problem, supervised learning with naïve architecture is more appropriate and gives more relevant results than an unsupervised learning model with the maximum weight spanning tree.

Actually, the superiority and performance of the naïve model in the study field of the failing companies have been confirmed in several previous studies, such as those of Suzaida. B. and

Muhammad. Z.A.H. (2012), Salehi and Abdini (2009), Sun and Shenoy (2007), Sarkar and Sriram (2001).

Conclusion

The objective of this study is to explain the causes of the Tunisian companies' default using one of the models of artificial intelligence, such as the Bayesian network. The application of the naive Bayesian network makes us conclude that the default probability of the Tunisian firms is positively correlated, on the one hand, with their general and long and medium term indebtedness levels, the size of their financial expenses and the levels of their personnel costs and, on the other hand, with the rate of their equipment renewal, their levels of general financial autonomy, their financial profitability ratios, the size of their depreciation allowances and their financial equilibrium ratios. Finally, we changed the learning algorithm by testing the maximum weight spanning tree (MWST) model the learning of which is unsupervised, unlike the naive Bayes model in which learning is supervised. The comparison of the results of both models makes us conclude that, in accordance with several previous studies, for a classification problem, the naive Bayesian network is more efficient than the MWST method since it gives a good ranking rate and a higher ROC index: Good ranking rate NBN = 91.28% > Good ranking rate MWST = 70.76% and ROC NBN index =84.96% > ROC MWST index =72.65%. In this study, we used the significance threshold of 1% for the selection of the variables and the exact inference method for the training of the model. Future studies may give different results if they take another significance threshold or if they use the approximate inference method with a Bayesian model

Appendix 1: Definition of the calculated financial ratios (initial list)

Ratios	Définition	Formula calculation
	Liquidity (and funding) ratios	
R1	Working capital ratio	Current assets / Current liability
R2	Ratio of reduced liquidity	Assets out of stock / current liabilitiy
R3	Immediate liquidity ration	Liquidity and equivalent / current liability
R4	(Fixed asset-hedging with permanent capital) working capital ratio	Permanent capital / non-current assets
R5	Cash ratio	Liquidity and equivalent of liquidity/ current bank account
R6	Funding structure	Overall debt / equity
R7	General financial autonomy	Equity capitals / complete balance sheet
R8	Long-term financial autonomy	(equity capital / permanent capital)
R9	Funding automy ratio in CT	Equity capital / total liabilities
R10	Debt ratio	Total liability/total asset
R11	Hedging rate of financial charges	Financial charges / turnover

R12	Net financial charges / operating income	Management (and rotation) ratio
The management ratios		
R13	Asset turnover ratio	Turnover/total assets
R14	Inventory turnover of tangible assets	Turnover / total fixed assets
R15	Equipment renewal ratio	Depreciation expense / fixed assets
R16	Importance rate of personnel costs	Personnel expenses/added value
R17	Collection of accounts receivable	Customers and associated account * 365) /sales
R18	Collection of accounts payable	Suppliers and associated account * 365) /purchases
R19	Importance of depreciation expenses	Depreciation expenses / added value
R20	Importance ratio of financial charges	Financial charges/added value
Profitability ratios		
R21	Financial return ratio	Net earnings /net equity capital
R22	Economic profitability ratio	Operating income / total assets
R23	Operational profitability ratio	Operating income/turover

The structure ratios		
R24	Ratio of fixed assets	Net capital / total assets
R 25	Weight of the receivables in the balance sheet structure	Total debt / total assets
R26	Weight of liquidity in the balance sheet structure	Liquidity and liquidity equivalent / total assets
R27	Equity capital ration	Equity capita/total liability
R28	Medium and long-term debt ratio	Medium and long-term debt/total liability
R29	Short-term debt ratio	Short-term debt/total liability
R30	Financial balance ratio	Working capital /total assets
R31	Evolution ratio of working capital	Working capital /turnover
R32	Size indicator	Log (total balance sheet)

Appendix 2: The significant partial correlations

Pair	Pearsons correlation	significance at 1% threshold
P (Y,R7 | R16) P (Y,R16 | R7)	-240 -0.805	0.006 0.000
P (Y,R7 | R19) P (Y,R19 | R7)	-0.307 -353	0.000 0.000
P (Y,R7 | R20) P (Y,R20 | R7)	-0.353 -0.229	0.000 0.009
P (Y,R10 | R16) P (Y,R16 | R10)	0.240 -0.805	0.006 0.000
P (Y,R10 | R19) P (Y,R19 | R10)	0.307 -0.353	0.000 0.000
P (Y,R10 | R20) P (Y,R20 | R10)	0.353 -0.229	0.000 0.009
P (Y,R15 | R19) P (Y,R19 | R15)	-0.898 -0.223	0.000 0.010
P (Y,R15 | R30) P (Y,R30 | R15)	-0.877 0.267	0.000 0.002
P (Y,R16 | R19) P (Y,R19 | R16)	-0.801 0.269	0.000 0.002
P (Y,R16 | R30) P (Y,R30 | R16)	-0.754 -0.286	0.000 0.001
P (Y,R19 | R21) P (Y,R21 | R19)	0.386 -0.246	0.000 0.005

P (Y,R19 | R28) P (Y,R28 | R19)	-0.386 0.246	0.000 0.005
P (Y,R19 | R30) P (Y,R30 | R19)	-0.254 -0.479	0.004 0.000
P (Y,R21 | R30) P (Y,R30 | R21)	-0.303 0.568	0.000 0.000
P (Y,R28 | R30) P (Y,R30 | R28)	0.303 -0.568	0.000 0.000

Appendix 3: Descriptive statistics of the explanatory variables

Descriptive statistics									
	Healthy companies					Failing companies			
Ratio	N	Min	Max	Moy	SD	Min	Max	Moy	SD
R7	195	-,7122	,8450	,4247	,1983	-6,8182	,9238	-,1424	,9474
R10	195	,1550	1,7122	,5753	,1983	,0762	7,8182	1,1424	,9474
R15	195	-,9769	,0000	-,6083	,1997	-,9639	,0000	-,4788	,2507
R16	195	-,3622	,0000	-,0867	,0730	-,6784	,0226	-,1346	,1352
R19	195	-,4267	,0000	-,0473	,0675	-3,2650	,0000	-,1301	,3171
R20	195	-2,1045	5,9841	-,0097	,4588	-5,4772	,0000	-,0991	,4211
R21	195	,1055	1,0000	,8697	,1827	,0904	1,0000	,7344	,2486
R28	195	,0000	,8945	,1303	,1827	,0000	,9096	,2656	,2486
R30	195	-,1665	,9318	,5064	,1907	-6,2573	,8321	-,2824	,8505

Appendix 4: discretization ratio

Nœuds 10					
Y	Discret	**Modalités**	**Agrégats**		
		0	0		
		1	1		
R7	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-4,238	-6.81820074	-4.237518025666667	Demandée : Égales distances - 3 Obtenue : Égales distances - 3
		<=-1,657	-4.237518025666667	-1.6568353113333334	
		>-1,657	-1.6568353113333334	0.923847403	
R10	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=2,657	0.076152597	2.6568353113333334	Demandée : Égales distances - 3 Obtenue : Égales distances - 3
		<=5,238	2.6568353113333334	5.237518025666667	
		>5,238	5.237518025666667	7.81820074	
R15	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-0,651	-0.976896114	-0.6512640759999999	Demandée : Égales distances - 3 Obtenue : Égales distances - 3
		<=-0,326	-0.6512640759999999	-0.32563203799999996	
		>-0,326	-0.32563203799999996	0.0	
R16	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-0,452	-0.678423237	-0.45228215799999993	Demandée : Égales distances - 3 Obtenue : Égales distances – 3
		<=-0,226	-0.45228215799999993	-0.22614107899999997	
		>-0,226	-0.22614107899999997	0.0	

R19	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-2,177	-3.265033547	-2.176689031333333	Demandée : Égales distances - 3 Obtenue : Égales distances – 3
		<=-1,088	-2.176689031333333	-1.0883445156666665	
		>-1,088	-1.0883445156666665	0.0	
R20	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-1,657	-5.477205588	-1.6567620586666671	Demandée : Égales distances - 3 Obtenue : Égales distances – 3
		<=2,164	-1.6567620586666671	2.163681470666666	
		>2,164	2.163681470666666	5.984125	
R21	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=0,394	0.09035432	0.39356954666666666	Demandée : Égales distances - 3 Obtenue : Égales distances – 3
		<=0,697	0.39356954666666666	0.6967847733333333	
		>0,697	0.6967847733333333	1.0	
R28	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=0,303	0.0	0.30321522666666667	Demandée : Égales distances - 3 Obtenue : Égales distances – 3
		<=0,606	0.30321522666666667	0.6064304533333333	
		>0,606	0.6064304533333333	0.90964568	
R30	Continu	**Modalités**	**Intervalles**		**Discrétisation**
		<=-3,861	-6.257266491	-3.860909628666667	Demandée : Égales distances - 3 Obtenue : Égales distances – 3

Bibliography

- **Arasti Z. (2011),** « An empirical study on the causes of business failure in Iranian context» African Journal of Business Management Vol. 5, pp. 7488-7498;
- **Ciampi F. et Gordini N. (2008),** *«Using Economic-Financial Ratios for Small Enteprises Default Prediction Modeling: an Empirical Analysis»*, in Oxford Business and Economics Conference Proceedings, Oxford, UK, pp 1-21;
- **Guilhot B. (2000)**, « *Défaillances d'entreprise : soixante-dix ans d'analyses théoriques et empiriques* », *Revue française de gestion*, pp.52-67 ;
- **Heckerman D., Geiger D., et Chickering M. (1994),** «*Learning Bayesian networks : The combination of knowledge and statistical data*», communication présentée dans la 10eme Conférence de l' uncertitude dans l'intelligence artificielle, pp 293–301;
- **Henrion M. (1990),** « *Propagating uncertainty in Bayesian networks by probabilistic logic sampling*», Kanal, L. N. & Lemmer, J. F. (Eds.), Uncertainty in Artifcial Intelligence, pp 149–163;
- **Lili S., et Prakash P.S., (2007) «***Using Bayesian Networks for Bankruptcy Prediction – Some Methodological Issues*», European Journal of Operational Research, vol 180, N°2, pp738-753;
- **Lounis H. (2006),** « Apprentissage Bayésien »,Notes de cours (Séminaire sur l'apprentissage automatique) » Programme de Doctorat en Informatique Cognitive, Université du Québec à Montréal ;
- **Mansouri Y., Chetouani B. et El Kihel, (2008),** «*Using Neural Networks for fault detection in a distillation column*», International Journal Computer Applications in Technology, vol. 32, pp. 181–186;
- **Murphy K. P. (2002),** «Active Learning of Causal Bayes Net Structure», Technical Report, Department of Computer Science, UC Berkeley;
- **Sarkar S., et Sriram R. S. (2001),** « *Bayesian Models for Early Warning of Bank Failures* », Management Science, vol 47, N°11, pp 1457–1475;
- **Salehi M. et Abedini, B. (2009), «***Financial Distress Prediction in Emerging Market: Empirical Evidences from Iran*». Journal of Contemporary Research in Business, Vol. 1, N°1, pp. 6-26;
- **Sun, L. & Shenoy, P. (2007)** "*Using bayesian networks for bankruptcy prediction*", European Journal of Operational Research, vol. 180, no. 2: 738-753;

- **Suzaida B. et Muhammad Z.A.H. (2012),** « *Validating a Bankruptcy Prediction by Using Naïve Bayesian Network Model: A case from Malaysian Firms*», International Conference on Economics, Business Innovation;
- **Ting J.A., D'souza A., et Schaal S. (2011),** « *Bayesian robot system identification with input and output noise*», Neural Networks, vol 24, pp 99–108;
- **Wainwright M.J. et Jordan M. (2008),** «*Graphical Models, Exponential Families, and Variational Inference*»,Foundations and Trends in Machine Learning Vol. 1, N° 1;
- **Yang J., Xu H., Pan L., Jia P., Long F., et Jie M. (2011),** « *Task scheduling using Bayesian optimization algorithm for heterogeneous computing environments*», Applied Soft Computing vol 11, pp 3297–3310;
- **Zammel M., et Khoufi W. (2016),** «*The Causes of Tunisian SME Failure*», Arabian J Bus Manag Review, vol 6, N°274.