Accounting and Management Information Systems
Vol. 14, No. 1, pp. 79-106, 2015

Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks

Aida Krichene Abdelmoula[1,a]

[a]Institut des Hautes Etudes Commerciales de Carthage,
University of Carthage, Tunisia

Abstract: Credit risk is defined as the risk that borrowers will fail to pay its loan obligations. In recent years, a large number of banks have developed sophisticated systems and models to help bankers in quantifying, aggregating and managing risk. The outputs of these models also play increasingly important roles in banks' risk management and performance measurement processes. In this study we try to tackle the question of default prediction of short term loans for a Tunisian commercial bank. We use a database of 924 credit records of Tunisian firms granted by a Tunisian commercial bank from 2003 to 2006. The K-Nearest Neighbor classifier algorithm was conducted and the results indicate that the best information set is relating to accrual and cash-flow and the good classification rate is in order of 88.63 % (for k=3). A curve ROC is plotted to assess the performance of the model. The result shows that the AUC (Area Under Curve) criterion is in order of 87.4% (for the first model), 95% (third model) and 95.6% for the best model with cash flow information.

Keywords: banking sector, risk assessment, default risk, k-Nearest-Neighbor classifier, ROC curve

JEL codes: B41, C14, C45, C53

1. Introduction

Bank credit risk assessment is widely used at banks around the world. Since, credit risk evaluation is very crucial, variety of techniques is used for risk level assessment. In addition, credit risk is one of the main functions of the banking

[1] *Corresponding author*: Department of Accounting, Carthage University
tel. (+216) 26.985.490; 1, rue de la Paix Ain Zaghouane, Tunis, TUNISIA, 2046
mail addresses: aidakrichene@yahoo.fr; aida.krichene@iscae.rnu.tn

community (Seval, 2008). Basel Committee on Banking Supervision defined credit risk as the potential of a bank borrower or counterparty will fail to pay its obligations in accordance with agreed terms (Okan veli safakli, 2007).

Banks classify clients according to their profile. While classifying, financial background of customers and subjective factors of customers are evaluated. Financial ratios play an important role for risk level calculation (Berk *et al.*, 2011). These ratios are objective and indicate the financial statement of business. Balance sheet, income statement and cash flows are some financial statements for collecting information to calculate objective financial ratios. There are many other subjective factors too; these depend on bank decision strategy and its mission according to (Berk *et al.*, 2011). The Basel Committee on Banking Supervision, in a consultative document, tried to provide banks and supervisors with guidance on sound credit risk assessment and valuation policies and practices for loans independently of the accounting framework applied. In this document the third principle states that "A bank's policies should appropriately address validation of any internal credit risk assessment models»[1].

The implementation of this principle turns out to be a daily decision based on a binary classification problem distinguishing good payers from bad payers (Karaa & Krichène, 2012). Surely, evaluating the insolvency plays an important role since a good estimate of the quality of a borrower can help to decide whether granting the requested loans or not. The Basel Committee recommends a choice between two broad methodologies for calculating their capital requirements for credit risk, either external mapping approach or internal rating system (Karaa & Krichène, 2012).

Although the external mapping approach is difficult to apply because of the unavailability of external rating grades, the internal rating is easy and simple to implement since a lot of techniques have been proposed in the literature to develop credit-risk assessment models. Additionally, the subprime crisis, which shakes down the American and European countries and shows the fragility of banking sector and cast some doubt on the accuracy and usefulness of agency ratings (Matoussi & Abdelmoula, 2009). In fact, Credit scoring methods are used to evaluate both objective and subjective factors. These techniques spread all around the world in 50's (Abramowicz *et al.*, 2003). By these methods, information collection from customer is formalized. Besides, the scoring system forms a basis for loan approval. These models include traditional statistical techniques such as logistic regression (Steenackers & Goovaerts, 1989), multivariate discriminant analysis (MDA) (Altman, 1968), classification trees (Davis *et al.*, 1992), neural network (NNs) models (Desai *et al.*, 1996, Matoussi & Abdelmoula, 2009; Karaa and Krichène, 2012) and nonparametric statistical models like k-nearest neighbour, Henley & Hand (1997). Recent contributions have proposed the employ of Bayesian classification rules using Naïve Bayes classifiers. The results of these studies demonstrated their frequent ability to do better than the most existing techniques. In this context, Sarkar and Sriram (2001), and Sun and Shenoy (2007) had been successfully applied to bankruptcy prediction.

In this research we try to tackle the following question: how banks can develop fairly accurate quantitative prediction models that can used as very early warning signals for default risk. The most of previous research look at business failure prediction from the mid-term and long-term prospects (failure vs non failure). In our paper, we examine the short-term prospect (payment vs. non payment of the short term loan at maturity). We try also to explore the case of a bank who wants to use prediction model to assess its credit risk (see failure prediction in Tunisia by (Matoussi *et al.*, 1999), financial distress prediction using Neural Networks by (Matoussi & Krichene, 2010; Abid & Zouari, 2000), financial distress in Egypt by (El-Shazly, 2002), credit scoring model for Turkey's micro & small enterprises by (Davutyan, 2006). In this study, we use a K-Nearest Neighbour classifier model to investigate the credit–risk.

This paper is organized as follow. In section 2 we provide the theoretical framework and Empirical Modelling supporting our research question our research design respectively. In section 3, we define data and methodology. In Section 4, we present our results and discussion. Finally, Section 5 concludes the paper and presents some limits.

2. Credit risk assessment of banks: theoretical framework and empirical modelling

2.1. Theoretical framework of credit risk problem: agency theory

One of the most important applications of agency theory to the lender-borrower problem is the derivation of the optimal form of the lending contract. In credit market, there is an information asymmetry between the borrower, who usually has better information about the investment project and its potential profits and risk, and the lender (the bank) who doesn't have enough and reliable information relating to investment project. This lack of information in quantity and quality is a source of problems before and after the transaction takes place. The presence of asymmetric information normally leads to moral hazard and adverse selection problems. This situation shows a classical principal-agent problem.

The principal-agent models of the agency theory may be divided into three classes according to the nature of information asymmetry (Karel, 2006). First, we find models with ex-post asymmetric information qualified as moral hazard. In this case, agent receives some private information after signing the contract. Moral hazard refers to a situation in which the asymmetric information problem is created after the transaction occurs. Since the borrower has relevant information about the project the lender doesn’t have, the lender runs the risk that the borrower will engage in activities that are undesirable from the lender’s point of view because they make it less likely that the loan will be paid back (Matoussi & Abdelmoula, 2009).

Second, we find models with ex-ante asymmetric information known as adverse selection models (Karel, 2006). In these models agent has private information already before signing the contract. Adverse Selection refers to a situation in which the borrower have significant information that the lender lack (or vice versa) about the quality of the project before the transaction takes place. This happens when the potential borrowers who are the most likely to produce an adverse outcome (bad credit risks) are the ones who are most active to get a loan and are thus most likely to be selected. In the simplest case, lenders’ price cannot differentiate between good and bad borrowers, because the riskiness of projects is unknown. Finally, we find the third class known as signalling models, in which the informed agent may divulge his private information through the signal which he sends to the principal (Karel, 2006).

This problem is traditionally considered in the framework of costly state verification, introduced by (Townsend, 1979). The essence of the model is that the agent, who has no endowment, borrows money from the principal to run a one-shot investment project. The agent is confronted with a moral hazard problem. Should he declare the true value or should he decrease the outcome of the project? This situation illustrates ex-post moral hazard. Moreover, we can also face a situation of ex-ante moral hazard, where the unobservable effort by agent during the project realization may impact the outcome of the project. Townsend (1979) indicated that the optimal contract which solves this problem is known as standard (or simple) debt contract. This standard debt contract is characterized by its face value, which should be repaid by the agent when the project is finished. As another theoretical justification for simple debt contract was considered by (Diamond, 1984), where the costly state verification was changed by a costly punishment. Hellwig (2000, 2001) indicated that the two models are equivalent only under the risk neutrality assumption. However, when we consider the introduction of risk aversion, the costly state verification model still working, but the costly punishment model does not survive.

In the real world, credit institutions can use either guarantee (collateral) or bankruptcy prediction modelling or both to face out the asymmetric information problem and its consequences on credit risk evaluation (Karaa & Krichène, 2012). We deal with this aspect in the next subsection.

2.2. Credit Risk Assessment and Bankruptcy Prediction: Related studies (works)

After the high number of profile bank failures in Asia, the regulators recognize the need and urge banks to employ advanced technology to assess the credit risk in their portfolios. Assessing the credit risk correctly also permits banks to engineer future lending transactions, so as to achieve targeted return/risk characteristics. The evaluation of credit risk needs the development of fairly accurate quantitative prediction models that can serve as very early warning signals for counterparty defaults.

Many researchers proposed two main approaches to deal with credit scoring in the literature. The first approach proposed by (Merton, 1974) and known as the structural or market based models where the default probability derivation is based on modelling the underlying dynamics of interest rates and firm characteristics. Initially, this approach is based on the asset value model, where the default process is endogenous, and relates to the capital structure of the firm. Default happens when the value of the firm's assets drops below some critical level (Crouhy *et al.*, 2000). The second approach is centered on the empirical or accounting based models where the relationship between default probability and characteristics of a firm is learned from the data instead of modelling this relationship. Raymond (2007), Thomas *et al.* (2002), Galindo and Tamayo (2000) synthesized some methods used in this context. In this regard we can cite the studies of Beaver (1966) and Altman (1968), bankruptcy prediction has been investigated intensively by academics and practitioners. Several models have been developed and tested empirically. Altman's popular Z-Score (Altman, 1968) is an illustration based on linear discriminant analysis, and was used to predict the probability of default of firms. Ohlsons O-Score (Ohlson, 1980) is based on generalized linear models or multiple logistic regression which have been used either to detect the best determinants of bankruptcy and the predictive accuracy rate of their occurrence. Neural network models were adapted and used in bankruptcy prediction (Atya, 2001; Matoussi & Abdelmoula, 2009). Their high power of prediction makes them a widely held alternative with the ability to integrate a very large number of features in an adaptive nonlinear model (Kay & Titterington, 2000).

A lot of researches have focused on the non-parametric methods class (e.g. k-nearest neighbor) (Henley & Hand, 1996), decision trees (Quinlan, 1992) and neural networks (Mcculloch & Pitts, 1943) have also been largely applied in the field of credit scoring. There are also some other approaches that combine several techniques to create a classification model such as Support Vector Machine (e.g. Lee and Chen, 2005; Lee *et al.*, 2002). West (2000) tried to compare the accuracy of credit scoring of five Artificial Neural Network models namely multilayer perceptron, radial basis function, fuzzy adaptive resonance, mixture-of-experts and learning vector quantization. In his study, West (2000) used two real world data sets Australian and German. He employed 10 fold cross validation for improving his predictive power. He indicated both good and bad credit rates. Finally, he compared the results against five other traditional techniques including linear discriminant analysis, logistic regression, k nearest neighbor, kernel density estimation and decision trees. The results indicate that the multilayer perception may not be the most accurate Artificial Neural Network model and that both the combination-of-experts and radial basis function Neural Network models should be considered for credit scoring applications. Also, between traditional methods, logistic regression is more accurate method and more precise than Neural Network models in average case.

According to Vera *et al.* (2012) "Despite the intense study of credit scoring, there is no consensus on the most appropriate classification technique to use." Baesens *et al.* (2003b) revealed that some conflicts can occur when comparing the findings of different studies. However Thomas *et al.* (2002) also suggested that most methods applied in credit scoring have similar levels of performance. In fact, for banks and financial institutions the reasons that may motivate the preference for a certain methods are the interpretability and the transparency (Martens *et al.*, 2009). According to Vera *et al.* (2002) "two aspects of methods for credit scoring are very important: that is the predictive performance, as well as the insights or interpretations that are revealed by the model."

2.3. Empirical research design

2.3.1. K-NN classifier algorithm

Banks are in a very competitive environment; thereby the service quality during credit risk assessment is very important. When customer demands for credit from bank, bank should evaluate credit demand as short as possible (Berk *et al.*, 2011) to gain competitive advantage. Additionally for each credit demand, the same process is repeated and constitutes a cost for the bank. Since the importance of credit risk analysis, most of techniques and models are developed by financial institution to decide whether to grant or not to grant credit (Çinko, 2006).

The classification methods can be classified into parametric and non-parametric problems. In fact, parametric methods are based upon the assumptions of normally distributed population and estimate the parameters of the distributions to solve the problem (Zhang *et al.,* 2007). However, according to Berry and Linoff (1997) non-parametric methods make no assumptions about the specific distributions involved, and are therefore distribution-free. The k-nearest neighbor classifier serves as an illustration of a non-parametric statistical approach. When given an unknown case, a K-NN classifier seeks the pattern space for the k training (Pranab & Radha, 2013) (cases that are similar to unknown cases. These k training cases are the K-nearest neighbors" of the unknown cases (Ravinder & Aggarwal, 2011).

K-NN classifier can be useful when the dependent variable takes more than two values such as high risk, medium risk and low risk. Moreover K-NN classifier requires an equal number of good and bad sample cases for better performance (Hand & Henley, 1997). According to Berry and Linoff (1997) "the choice of ***k*** also affects the performance of the k-NN algorithm. This can be determined experimentally. Starting with k=1, we use a test case to estimate the error rate of the classifier. This process is repeated each time by incrementing k to allow for one more neighbors. The K-value that gives the minimum error rate may be selected. In general, larger the number of training samples is, the larger the value of k will be."

2.3.2. ROC curve as a classifier performance

A Receiver Operating Characteristics (ROC) is a generally useful performance graphing method. In other word, ROC graph is a method for visualizing, organizing and selecting classifiers based on their performance Fawcett (2006). Spackman (1989) was the earliest adopters of ROC graphs in machine learning. He demonstrated the value of ROC curves in evaluating and comparing algorithms (Fawcett, 2006). In fact, the use of ROC graphs in the machine learning community has increased in recent years. Since that simple classification accuracy is often a poor metric for measuring performance (Provost & Fawcett, 1997; Provost *et al.*, 1998). Besides, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs (Fawcett, 2006).

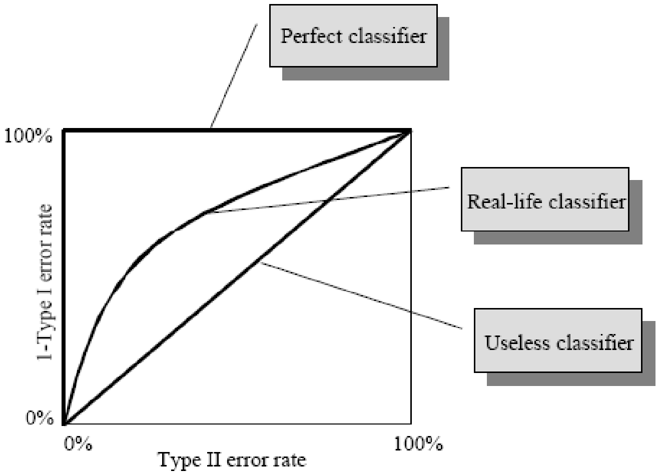


Figure1. An example of ROC curve adapted from Yang (2002: 18)

2.3.3. The criterion of the area under a curve ROC (AUC)

A ROC curve is a two-dimensional representation of classifier performance. According to Fawcett (2006), "to compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance". To do so, many researchers such as Bradley (1997) and Hanley and McNeil (1982) recommend the use of a common method which is to calculate the area under the ROC curve, abbreviated AUC. The AUC is defined as a portion of the area of the unit square; its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5 (Fawcett, 2006).

3. Methodology

The need for models that predict defaults correctly is very important, because in commercial bank the credit risk measurement is crucial for each client to discriminate reliable clients from not reliable. Among the quantitative methods for solving credit risk evaluation problems, the simple Bayesian classifier was applied for estimating the posterior probabilities of default. In fact, Antonakis and Sfakianakis (2009) showed that the posterior probability of an event is the probability of an event after collecting some empirical data. Rosner (2006) demonstrated that the posterior probability is obtained by integrating information from the prior probability with additional data related to the event in question. According to Mileris (2010) "often analysis begins with initial or prior probability estimates for specific events of interest. Then from sources such as a sample we obtain additional information about the events. Given this new information the prior probability values can be updated by calculating revised probabilities, referred to as posterior probabilities". Anderson *et al.* (2007) demonstrated that Bayesian theorem provides a means for making these probability calculations.

In our research we use a sample of the bank credit files composed of 924 files of short term loan granted to Tunisian companies from 2003 to 2006.

3.1. Sample and data

Let's recall that our objective is to use k-NN[2]classifier methodology for default prediction of bank's commercial loans. But, in order to solve a problem using k-NN algorithm, we have to collect data for training purposes. The training data set includes a number of cases, each containing values for a range of input and output variables. The first decision we need to make is which variables to use. The second one concerns the subjects we want to predict their behaviour. For our case the variables are indicators of default risk and the subjects are borrowers. The data collected for our investigation came from a large private commercial bank (BIAT)[3]. We choose a private bank in order to avoid the potential inefficiency of public banking sector, whose decision is sometimes dictated by government choices. Our choice to work with short-term commercial credits is motivated by the fact that this type of credits represents the major part of loans granted by commercial bank and is subject to renewal every year. In fact, loans with maturities of one year or less comprise more than half of all commercial bank loans (Revsine *et al.,* 1999). For the case of the BIAT this ratio was around 40% during 2006 and 2007.

In our investigation, we use a database of 924 files granted to industrial Tunisian companies by a commercial bank from 2003 to 2006. This period was chosen because it corresponds to a central bank instruction, in which it asks bank to provide credit risk classes for their borrowers. For the case the BIAT, by the end of every quarter, it classifies these files into five clusters, each one corresponding to a risk class. Files without delay of payment correspond to the healthy firms. The four remaining classes correspond to four riskier classes of three, six, nine and one year (or more) delay of payment respectively. We group these four classes in one class (risky companies).

3.2. Variables measurement

3.2.1. Dependent variable

In this research, we try to study the probability of default. The dependent variable is a dummy variable, which equals 0 if the classified is healthy and 1 if the classified is risky. Hence:

Y = 0 if no delay of payment (healthy)
Y = 1 if more there is more than 3 month delay (risky)

3.2.2. Independent variables

Default risk prediction depends on a good evaluation of the couple risk-return of a company. Financial ratios, commonly used, are calculated from financial statements (balance sheet, income and cash flow statement). Financial ratio analysis classifies the ratios into groups which states about different facets of a company's finances and operations (liquidity, activity or operational, leverage and profitability).

In our experiment we keep the same variables used in the study of (Karaa & Krichène, 2012). So the database is composed of 24 financial and non-financial indicators. The financial indicators are inspired from Altman’s popular Z-Score and recommended textbooks in financial statement analysis and valuation (Berstein

& Wild, 1998; Revsine *et al.*, 1999; and Palepu *et al.*, 2000). The financial indicators measure liquidity (working capital, operating activity and cash flow), Leverage, long term solvency and Profitability. The non-financial variables used in this research are firm size and collateral (Karaa & Krichène, 2012).

Table 1. Variables definition and measure[4]

RISK FACET	CODE	VARIABLE DEFINITION	VARIABLE MEASURE
Liquidity indicators	R1	Long term financing of Working capital	(Shareholders'equity+non current liabilities)- non current assets
	R2	Working capital requirement	$\frac{\text{Working capital}}{\text{(Shareholders'equity+non current liabilities)-non current assets}}$
	R3t	Account receivable liquidity	$\frac{\text{Provision for doubtful accounts}}{\text{Gross account receivables}}$
	R4	Current ratio	$\frac{\text{Current assets}}{\text{Current liabilities}}$
	R5	Quick ratio	$\frac{\text{Current assets-inventories}}{\text{Current liabilities}}$
	R6	Cash flow ratio	$\frac{\text{Operating cash flow}}{\text{Current liabilities}}$
	R7	Inventory turnover	$\frac{\text{Sales}}{\text{inventories}}$
Leverage and solvency indicators	R8	Debt Cash Flow Coverage Ratio	$\frac{\text{Cash flow}}{\text{Total debts}} = \frac{\text{Net income +depreciation}}{\text{Total debts}}$
	R9	Liabilities to equity ratio	$\frac{\text{Total liabilities}}{\text{Shareholders' equity}}$
	R10	Net debt to equity ratio	$\frac{\text{Short term debt+long term debt – cash and marketable securities}}{\text{Shareholders' equity}}$
	R11	Debt to capital ratio	$\frac{\text{Short term debt+long term debt}}{\text{Short term debt+long term debt + Shareholders'equity}}$
	R12	Long term debt to assets	$\frac{\text{Long term debt}}{\text{Total assets}}$

	R13	Long term debt to tangible assets	$\frac{\text{Long term debt}}{\text{Total tangible assets}}$
	R14	Interest coverage ratio	$\frac{\text{Operating income before taxes and interest}}{\text{Interest expense}}$
Profitability indicators	R15	Net profit margin	$\frac{\text{Net income}}{\text{Total operating revenue}}$
	R16	Gross profit margin	$\frac{\text{Earnings before interest and taxes}}{\text{Total operating revenue}}$
	R17	Return on invested capital	$\frac{\text{Net income}}{\text{Total assets}}$
	R18	Return On Equity (ROE)	$\frac{\text{Net income}}{\text{Stockholders equity}}$
Ratios used by the bank	R19	Fixed asset to debt ratio	$\frac{\text{Net fixed assets}}{\text{Total debt}}$
	R20	Short term debt to sales ratio	$\frac{\text{Short term debt}}{\text{Total sales}}$
	R21	Financial expenses to revenue ratio	$\frac{\text{Financial expenses}}{\text{Total revenue}}$
	R22	Fixed asset turnover	$\frac{\text{Sales}}{\text{Fixed assets}}$
Other variables	V01	collateral	LOG(GUARANTEE)
	V02	Firm size	LOG(TOTAL ASSETS)

4. Empirical results

4.1. Descriptive analysis

To get an insight about our data before performing the k-NN classifier models, we will achieve a test of mean differences between the two risks classes defined above (table 2). The summary statistics and the mean differences can be seen as an analysis similar to Beaver (1963). In Table 2 we expose the descriptive statistics of our data. When we run mean differences analysis between the two risks classes (healthy and risky groups). Such analysis allows us to verify if there is a difference between the two classes in terms of financial ratios. Table 2 presents some summary statistics for the two risks classes.

Table 2. Group means

Ratios	Code	Mean	Std. Deviation
R2:	,00	***16,8191***	58,85241
	1,00	***8,5990***	15,69326
R3:	,00	,0471	,13891
	1,00	,0568	,14135
R4:	,00	***2,9623***	7,05638
	1,00	***3,2328***	8,05572
R6:	,00	***2,0391***	22,41881
	1,00	***-,6450***	38,61664
R7:	,00	,0439	,10179
	1,00	,0757	,14164
R8:	,00	***1,4900***	2,00318
	1,00	***1,0742***	,91636
R10:	,00	,0452	,33929
	1,00	,0347	,16519
R11:	,00	***,0569***	,15137
	1,00	***,0166***	,10049
R12 :	,00	***,4993***	2,33091
	1,00	***,2348***	1,14838
R13:	,00	,7708	,97464
	1,00	,7151	,58013
R14:	,00	***,2274***	1,06959
	1,00	***,0588***	,74966
R15:	,00	***5,0372***	55,16137
	1,00	***13,4255***	16,99936
R18:	,00	***,6227***	2,93441
	1,00	***7,4529***	54,44136
R19:	,00	1,8072	2,80796
	1,00	1,7822	3,89486
R20:	,00	1,1982	2,38237
	1,00	1,1215	3,14473
R21:	,00	***,0634***	,30944
	1,00	***,2492***	2,91846
R22:	,00	***,0115***	,43979
	1,00	***-,0284***	,25000

00: corresponds to healthy group
01: corresponds to risky group

Table 2 presents significant mean differences between the two groups for some ratios (R_2; R_4; R_6; R_8 R_{11}; R_{12}; $R_{14;}$ R_{15}; R_{18}; $R_{21}$and R_{22}) and no significant differences for others (R_1; R_3; R_5; R_7; R_9; R_{10}; R_{11}; R_{13}; R_{16}; R_{17}; R_{19} and R_{20}). Globally, they tell us that the liquidity risk does not differentiate the two groups (Karaa & Krichène, 2012). The leverage and solvency ratios do better in discriminating the two groups. For others indicators (coverage and profitability), the results are mitigated. For example, while return on equity (R_{18}) shows a significant difference gross profit margin (R_{16}) and return on invested capital (R_{17}) are not.

When we look at the significance of mean differences, we notice that globally the good indicators are superior in the healthy group, while the bad indicators are higher in the risky group. For example the mean of cash flow ratios (R_6), Working capital requirement (R_2), leverage and solvency ratios (R_{11}, R_{12}, R_{14} and R_8) is bigger in health group. Current ratio (R_4), profitability ratios (R_{18} and R15), have a higher mean in the risky group.

4.2. Results and discussion

In our experiment, we build up three types of K-NN classifier. The first classifier uses data on financial ratios (cash-flows excluded). It will be referred as 'Non cash-flow model'. The second model uses data on all ratios indicators (cash-flows included, collateral excluded). It will be referred as 'Cash-flow model'. The third model uses all indicators of the study. It will be referred as 'full information model'.

According to Rafiul *et al.* (2008) "the *k*-nearest neighbor (*k*-NN) technique, due to its interpretable nature, is a simple and very intuitively appealing method to address classification problems. However, choosing an appropriate distance function for *k*-NN can be challenging and an inferior choice can make the classifier highly vulnerable to noise in the data". In our investigation, we tested using different values of *k* (2, 3, 4 and 5). Based on this testing, for *k*-NN we identified the best value of *k* which produced the best classification performance and this is what is reported in the result tables 3, 4 and 5.

Table 3. Results for Non Cash-Flow models (Appendix 1)

Panel 1: k-NN classifier with variation of the parameter k=2 (appendix 1 panel 1)

	K=2	
	Healthy	**Risky**
Healthy companies	**358**	**100**
Risky companies	**100**	**366**
% Total Good and Bad	Classification	
Good classification	78.35%	
Bad classification	21.64%	

Panel 2: K-NN with k=3 (appendix 1 panel 2)

	K=3	
	Healthy	**Risky**
Healthy companies	364	94
Risky companies	78	388
% Total Good and Bad	**Classification**	
Good classification	**81.38%**	
Bad classification	18.62%	

Panel 3: K-NN with k=4 (Appendix 1 panel 3)

	K=4	
	Healthy	**Risky**
Healthy companies	**332**	**126**
Risky companies	**122**	**344**
% Total Good and Bad	Classification	
Good classification	73.16%	
Bad classification	26.84%	

Panel 4: K-NN with k=5 (Appendix 1 panel 4)

	K=5	
	Healthy	**Risky**
Healthy companies	**331**	**127**
Risky companies	**124**	**342**
% Total Good and Bad	Classification	
Good classification	**72.83%**	
Bad classification	**27.16%**	

We can see from these results (panel 1, 2, 3 and 4) that the global good classification rate is getting better when we fixed the number of the parameter k to 3. In fact, the good classification rate is in order of 81.38% for the best model with k=3 for the other models with k=2, 4 and 5 the good classification rate is respectively of 78.35%, 73.16% and 72.83%. A lot of researches have examined the criterion of type I and II errors. According to Yang (2002) «Type I error rate is also called a rate or credit risk, it is the rate of 'bad' customers being categorized as 'good'. When this happens, the miss-classified 'bad' customers will become default. Therefore, if a credit institution has a high a rate, which means the credit granting policy, is too generous, the institution is exposed to credit risk».

Type II error rate is called also a commercial risk; it is the rate of 'good' client being classified as 'bad'. When this happens, the miss-classified 'good' client are rejected, the bank supports (endure) therefore an opportunity cost caused by the loss of 'good' customers. Bogess (1967) showed that if a credit institution has a high type II error for a long period, which means it takes a long time restrictive credit granting policy Yang (2002), it may lose its share in the market. The credit institution is therefore exposed to commercial risk.

Table 4. Results for Cash-Flow models (Appendix 2 panels 1,2,3 and 4)

	K=2		K=3		K=4		K=5	
	Healthy	**Risky**	**Healthy**	**Risky**	**Healthy**	**Risky**	**Healthy**	**Risky**
Healthy companies	**395**	**63**	**409**	**49**	**387**	**71**	**375**	**83**
Risky companies	**59**	**407**	**56**	**410**	**72**	**394**	**92**	**374**
% Total Good and Bad Classification								
Good classification	**86.79%**		**88.63%**		**84.52%**		**81.06%**	
Bad classification	**13.20%**		**11.37%**		**15.48%**		**19.94%**	

Table 5. Results for full information models (Appendix3 panels 1, 2, 3 and 4)

	K=2		K=3		K=4		K=5	
	Healthy	**Risky**	**Healthy**	**Risky**	**Healthy**	**Risky**	**Healthy**	**Risky**
Healthy companies	**393**	**65**	**406**	**52**	**381**	**77**	**383**	**75**
Risky companies	**69**	**397**	**69**	**397**	**99**	**367**	**113**	**353**
% Total Good and Bad Classification								
Good classification	**85.5%**		**86.90%**		**80.95%**		**79.65%**	
Bad classification	**14.50%**		**13.10%**		**19.05%**		**20.35%**	

The classification results for the two models (cash flow and full information models) are presented in Table 4 and 5. The best performances among that of the reported classifiers are marked in bold and red.

From tables 4 and 5 we can see that the best model which shows the best classification rate is the one associating accrual and cash flow information (table 4) with a good classification rate of 88.63% versus 86.90% for the third model with full information. Let's recall that our objective is to find the class label for the new point. The algorithm has different behavior based on k and in this research we choose the value of K. We can also conclude that the best parameter k –NN is set to 3 for all models in this research.

The variation of the parameter k to 3 has improved the results. The good classification rate is getting better. Moreover, the model has reduced the error type

I from 16.73% to 12% (Table 6) and the error type II is reduced from 20.52% to 10.69% when we introduce cash flow information.

Table 6. Criterion of the type I and II error

	ERROR	K=2	K=3	K=4	K=5
NON CASH FLOW MODEL	Type I	**21.83%**	**16.73%**	**27.51%**	**27.25%**
	Type II	**21.45%**	**20.52%**	**26.18%**	**27.72%**
CASH FLOW MODEL	Type I	**12.66%**	**12.01%**	**15.45%**	**19.74%**
	Type II	**13.75%**	**10.69%**	**15.50%**	**18.12%**
FULL INFORMATION MODEL	Type I	**14.8%**	**14.80%**	**21.24%**	**24.24%**
	Type II	**14.19%**	**11.35%**	**16.81%**	**16.37%**

In this research, we would like to assess credit risk using a selection of financial ratio recommended in debt contracts. The predictions on the selection of financial ratio illustrate relation between financial ratios and credit risk. This evidence is well known in the practitioner and academic literature (Demerjian, 2007). In fact, textbooks emphasize the role of ratios in evaluating credit quality (Lundholm & Sloan, 2004), while academic studies conclude that financial ratios serve to provide signals about borrower credit risk when used as covenants (Smith & Warner, 1979; Dichev & Skinner, 2002).

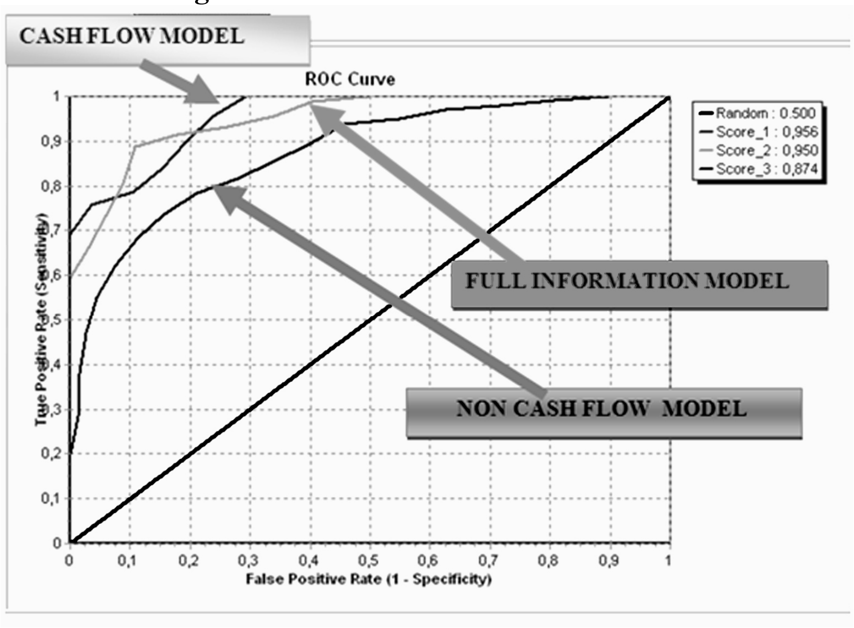
4.3. The ROC curve

A ROC curve for the perfect classifier, which orders all 'bad' cases before 'good' cases, is the curve follows the two axes. It would classify 100% 'bad' cases into class 'bad' and 0% 'good' cases into class 'bad' for some value of the sill. According to Yang (2002) “a classifier with a ROC curve which follows the 45° line would be useless. It would classify the same proportion of the 'bad' cases and 'good' cases into the class 'bad' at each value of the threshold; it would not separate the classes at all. Real-life classifiers produce ROC curves which lie between these two extremes”.

To evaluate the performance of the curve we have to use a measure given by the Area under the ROC Curve (denoted as AUC) (Hand, 1997). The curve that has a larger AUC is better than the one that has a smaller AUC.

We can note that the criterion of AUC is of the order of 95.6% for the best model (cash flow model). This score is larger than 50% and it is a good score. This result confirms the good classification rate found in the previous section. We can conclude that cash flow information is a good indicator for bankers who want to evaluate credit applicant.

Figure 2. ROC curve of three models



5. Conclusion

Commercial banks that grant client borrower loans need consistent models that can correctly detect and predict defaults. Moonasar (2007) emphasized that the one of the fundamental tasks which any bank has to deal with, in the current competitive and turbulent business environment, is to reduce its credit risk. Traditionally, we employ scoring methods to estimate the credit worthiness of a credit applicant. In fact, the quantitative method known as credit scoring has been developed for the credit assessment problem (Yang, 2002). Credit scoring is basically an application of classification methods, which classify borrower into different risk groups. The objective of Scoring methods is to predict the probability that a borrower or counterparty will default (Komor´ad, 2002). In credit risk evaluation, Credit scoring is a key methods, that help financial institution to make decision whether or not to grant credit to customer Thomas (2002). According to Moonasar (2007) “a common approach of credit scoring is to apply a classification technique on data of previous customers (both good credit customers and delinquent customers) in order to find a relationship between the customers characteristics and potential failure to service their debt. Institutions use credit scoring techniques (utilizing information from the consumers past credit history and current economic conditions) to determine which applicants will pay back their liabilities". An accurate classifier is necessary to differentiate between new potential good and bad credit applicant.

In this article we evaluate the credit risk for a Tunisian bank through modelling the default risk of its short term loans. We used a data base of 924 credit files from 2003 to 2006. In our evaluation, a K- Nearest Neighbor classifier algorithm was conducted and we tested using different values of k (2, 3,4 and 5). The criterion used for assessing performance is the minimization of the bad risk rate. We build up three types of K-NN classifier:

❶ The first classifier is non cash-flows model

❷ The second classifier is cash-flows model

❸ The third classifier is full information model

The main results show that the best K-NN with k=3 for the three models, and the best global classification rate is in order of 88.63% (second classifier). Moreover, to evaluate the performance of the model curve ROC is plotted. The result shows that the AUC (**A**rea **U**nder **C**urve) criterion is in order of 95.6%. Our study is, however, incomplete in the sense that it didn't show how one can use these results in the implementation of the Basel II or III accord in Tunisia.

References

Abid, F. & A. Zouari (2000) "Financial distress prediction using neural networks", http://ssrn.com/abstract=355980 or DOI: 10.2139/ssrn.355980.

Abramowicz, W. M. Nowak, J. Sztykiel (2003) "Bayesian networks as a decision support tool in credit scoring domain", Idea Group Publishing

Altman, E. I. (1968) "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *Journal of Finance* , vol. 23: 589–609

Anderson, D.R., Sweeney, D.J., Freeman, J., Williams T.A. & Shoesmith, E. (2007) "Statistics for business and economics", London: *Thomson Learning EMEA*

Antonakis, A. C. & Sfakianakis, M. E. (2009) "Assessing naive bayes as a method for screening credit applicants", *Journal of Applied Statistics*, vol. 36: 537-545

Atiya, A.F. (2001) "Bankruptcy prediction for credit risk using neural nets: a survey and new results", *IEEE Transactions on Neural Nets*, vol. 12 (4): 929-935

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. & Vanthienen, J. (2003) "Benchmarking state-of-the-art classification algorithms for credit scoring", *Journal of the Operational Research Society*, vol. 54 (6): 627–635

Beaver, W. (1966) "Financial ratios as predictors of failure. Empirical research in accounting: Selected studies", *Journal of Accounting Research,* vol. 5: 71–111

Berk Bekiroglu Hidayet Takci1 & Utku Can Ekinci (2011) " Bank credit risk analysis with bayesian network decision tool" *International Journal Of Advanced Engineering Sciences And Technologies,* vol. 9, no. 2: 273-279

Berry. M.J.A. & Linoff, G.S (1997) *Data mining techniques for marketing, sales, and customer support*, John Wiley & Sons, Inc.

Berstein, L. A. & Wild J.J. (1998) *Financial statement analysis: theory, application, and interpretation*, sixth Edition, McGraw-Hill

Bogess, W. P. (1967) "Screen -test your credit risks", *Harvard Business Review*, vol. 45, no. 6: 113-122

Bradley, A.P. (1997) "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recogn*, vol. 30(7): 1145-1159

Çinko, M. (2006) "Comparison of credit scoring tecniques: Istanbul ticaret üniversitesi sosyal bilimler", *Dergisi,* vol. 9: 143-153

Crouhy, M.; Galai, D.; Mark, R. (2000) "A comparative analysis of current credit risk models", *Journal of Banking and Finance,* vol. 24, no. 1: 59-117

Davis R. H., Edelman, D.B. & Gammerman, A.J. (1992) "Machine learning algorithms for credit-card applications", *IMA Journal of Management Mathematics*, vol. 4: 43-51

Davutyan, N. & Özar, S. (2006) "A credit scoring model for Turkey's micro & small enterprises (MSE's)," *13th Annual ERF Conference, Kuwait*, 16 – 18 December 2006

Demerjian, P. R. W (2007) "Financial ratios and credit risk: the selection of financial ratio covenants in debt contracts", working paper, workshop Stephen M. Ross School of Business University of Michigan, January 11

Desai, V. S., Crook, J. N. & Overstreet, G. A. (1996) "A comparison of neural networks and linear scoring models in the credit union environment", *European Journal of Operational Research*, vol. 95(1): 24–37

Diamond, D.W (1984) "Financial intermediation and delegated monitoring", *Review of Economic Studies*, vol. 51: 393–414

Dichev, I. & Skinner, D. (2002) "Large-sample evidence on the debt covenant hypothesis", *Journal of Accounting Research*, vol. 40 (4): 1091-1123

El-Shazly, A. (2002) "Financial distress and early warning signals: a non-parametric approach with application to Egypt", *9th Annual ERF Conference*, Emirates, October 2002

Fawcett, T. (2006) "Roc analysis in pattern recognition", *Pattern Recognition Letters*, vol. 27; no. 8: 861-874

Galindo, J. & Tamayo, P. (2000) “Credit risk assessment using statistical and MachineLearning: basic methodology and risk modeling applications”, *Computational Econ*omics, vol. 15(1-2): 107- 143

Hand, D. J. (1997) *Construction and assessment of classification rules*, Wiley series in probability and statistics, John Wiley & Sons

Hand, J. & Henley, W. (1997) “Statistical classification methods in consumer credit scoring”, *Computer Journal of the Royal Statistical Society Series a Statistics in Society*”, vol. 160, no. 3: 523-541

Hanley, J.A. & McNeil, B.J. (1982) “The meaning and use of the area under a receiver operating characteristic (ROC) curve”, *Radiology*, vol. 143: 29–36

Hellwig, M. (2000) “Financial intermediation with risk aversion”, *Review of Economic Studies*, vol. 67(4): 719–742

Hellwig M. (2001) “Risk aversion and incentive compatibility with ex post information Asymmetry”, *Economic Theory*, vol. 18 (2):415–438.

Henley, W.E. & Hand, D.J. (1997) “Statistical classification methods in consumer credit scoring: a review”, *Journal of the Royal Statistical Society. Series A* (Statistics in society), vol. 160, no. 3: 523- 541

Henley, W. E. & Hand, D. J. (1996) “A k-Nearest-Neighbour classifier for assessing consumer credit risk”, *The Statistician*, vol. 45(1): 77

Karaa,A. & Krichène, A. (2012) “Credit–risk assessment using support vectors machine and multilayer neural network models: a comparative study case of a Tunisian bank”, *Accounting and Management Information Systems,* vol. 11, no. 4: 587–620

Karel.J. (2006) “Agency theory approach to the contracting between lender and borrower” *Acta Oeconomica Pragensia,* 14/3

Kay, J. & Titterington, M. (eds) (2000) “*Statistics and Neural Nets, Advances at the Interface”*, Oxford University Press

Komorad, K. (2002) “On credit scoring estimation”, Institute for statistics and econometrics, Humboldt University, Berlin

Lee, T. & Chen, I. (2005) “A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines”, *Expert Systems with Applications*, vol. 28(4): 743–752

Lee, T., Chiu, C., Lu, C. & Chen, I. (2002) “Credit scoring using the hybrid neural discriminant technique”, *Expert Systems with Applications*, vol. 23(3): 245-254

Lundholm, R. & Sloan, R. (2004) *Equity valuation & analysis*, New York; McGraw-Hill/Irwin

Martens, D., Van Gestel, T., De Backer, M., Haesen, R., Vanthienen, J. & Baesens, B. (2009) “Credit rating prediction using ant colony optimization”, *Journal of the Operational Research Society*, vol. 61(4): 561–573

Matoussi, H. & Abdelmoula,A. (2009) “Using a neural network-based methodology for credit–risk evaluation of a Tunisian bank”, *Middle Eastern Finance and Economics* Issue 4

Matoussi, H. & Krichène Abdelmoula, A. (2010) "Credit risk evaluation of a Tunisian commercial: Bank: logistic regression versus Neural Network Modelling", *Accounting and Management Information Systems*, vol. 9, no. 1

Matoussi, H., Mouelhi, R. & Salah, S. (1999) "La prédiction de faillite des entreprises tunisiennes par la régression logistique", *Revue Tunisienne des Sciences de Gestion*, vol. 1: 90-106

Mcculloch, W. & Pitts, W. (1943) "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysic*, vol. 5: 115-133

Merton, R. (1974) "On the pricing of corporate debt: The risk structure of interest rates," *Journal of Finance*, vol. 29: 449-470

Mileris, R. (2010) "Estimation of loan applicants default probability applying discriminant analysis and simple bayesian classifier", *Economics and Management*, vol. 15: 1078-1084

Moonasar, V. (2007) "Credit risk analysis using artificial intelligence: evidence from a leading South African banking institution", *Research Report: Mbl3*

Ohlson, J. A. (1980) "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, vol. 18: 109-131

Okan veli şafakli (2007) "Credit risk assessment for the banking sector of Northern Cyprus", Banks and Bank Systems, vol. 2

Palepu K.G., Healy, P.M. & Bernard, V.L. (2000) *Business analysis & valuation using financial Statements*, second Edition, South – Western College Publishing

Pranab Kumar D. G., Radha Krishna, P., (2013) «Database management system oracle SQL AND PL/SQL" *PHI Learning Pvt. Ltd.,* 576 pages

Provost, F. & Fawcett, T. (1997) "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions", In: *Proc. Third Internat. Conf. on Knowledge Discovery and Data Mining (KDD-97). AAAI Press, Menlo Park*

Provost, F., Fawcett, T. & Kohavi, R. (1998) "The case against accuracy estimation for comparing induction algorithms", In: Shavlik, J. (Ed.) *Proc. ICML-98. Morgan* Kaufmann, San Francisco, Available from: <http://www.purl.org/NET/tfawcett/papers/ICML98-final.ps.gz>.

Quinlan, J. R. (1992) *C4.5 "programs for machine learning"*, Morgan Kaufmann Publishers Inc., California

Rafiul, H., Marufhossain, M., Bailey, J. & Kotagiri Ramamohanarao (2008) "Improving k-Nearest Neighbour classification with distance functions based on receiver operating characteristics", *Proceeding ECML PKDD '08 Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases* - Part I, pp. 489-504, Springer-Verlag Berlin Heidelberg

Ravinder, S. & Aggarwal, R.R. (2011) "Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data", *International Journal of Computer Theory and Engineering*, vol. 3, no. 5

Raymond, A. (2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press, United States of America, 1st edition.

Revsine, L., Collins, D.W. & Johnson, W.B. (1999) *Financial Statement and Analysis*, Prentice Hall, New Jersey.

Rosner, B.A. (2006) *Fundamental of Biostatistics*, Taunton: Quebecor World

Rumelhart, D.E., Hinton, G.E. & McClelland, J.L. (1986) "A general framework for parallel distributed processing", In "Parallel Distributed Processing: explorations in the microstructure of cognition", vol. 1, pp. 45-75

Sarkar, S. & Sriram, R.S. (2001) "Bayesian models for early warning of bank failures", *Management Science*, vol. 47(11): 1457-1475

Seval, S. (2008) "Credit risk and Basel II", Credit Risk Solutions Inforsense

Smith, C. & Warner, J. (1979) "On financial contracting", *Journal of Financial Economics,* vol. 7: 117-161

Spackman, K.A. (1989) "Signal detection theory: Valuable tools for evaluating inductive learning", In: Proc. Sixth Internat. Workshop on Machine Learning, Morgan Kaufman, San Mateo, CA, pp. 160-163.

Steenackers, A. & Goovaerts, M.J. (1989) "A credit scoring model for personal loans", *Insurance Mathematics and Economics*, vol. 8: 31-34

Sun, L. & Shenoy, P. (2007) "Using bayesian networks for bankruptcy prediction", *European Journal of Operational Research*, vol. 180, no. 2: 738-753

Thomas, L. C., Edelman, D. B. & Crook, J. N. (2002) *Credit scoring & its applications*, *Society for Industrial Mathematics*, Philadelphia, 1st edition.

Thomas, L.C. (2002) "A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers", *International Journal of Forecasting*, vol. 15: 149-172

Townsend, R. M. (1979) "Optimal contracts and competitive markets with costly state verification", *Journal of Economic Theory*, vol.21 (2): 265-293

Vera L. M., Dries F.B. & Van den Poel, D. (2012) "Enhanced decision support in credit scoring using bayesian binary quantile regression", Working Paper August

West, D. (2000) "Neural network credit scoring", *Computer & Operations Research*, vol. 27 (11): 1131-1152

West, D. (2000) "Neural network credit scoring model", *Computational Operation Research* vol. 27: 1131-1152

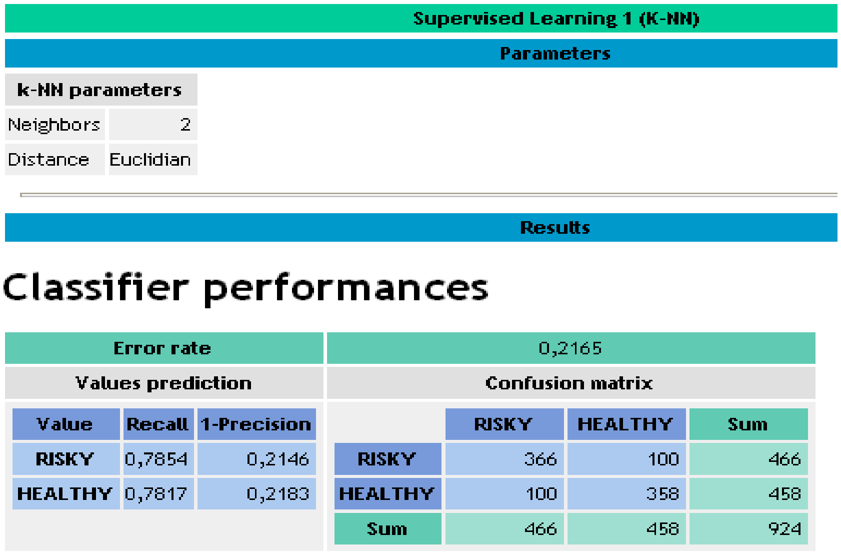
Yang, L. (2002) "The evaluation of classification models for credit scoring", Working Paper no. 02/2002 Edit. Matthias Schumann University of Göttingen Institute of computer science

Zhang, D., Huang, H., Chen, Q. & Jiang, Y. (2007) "Comparison of credit scoring models", *Third international conference of Natural Computation*, vol. 1

APPENDIX

APPENDIX 1: NON CASH FLOW MODEL

Panel 1: K-NN with k=2



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	2
Distance	Euclidian

Results

Classifier performances

Error rate	0,2165

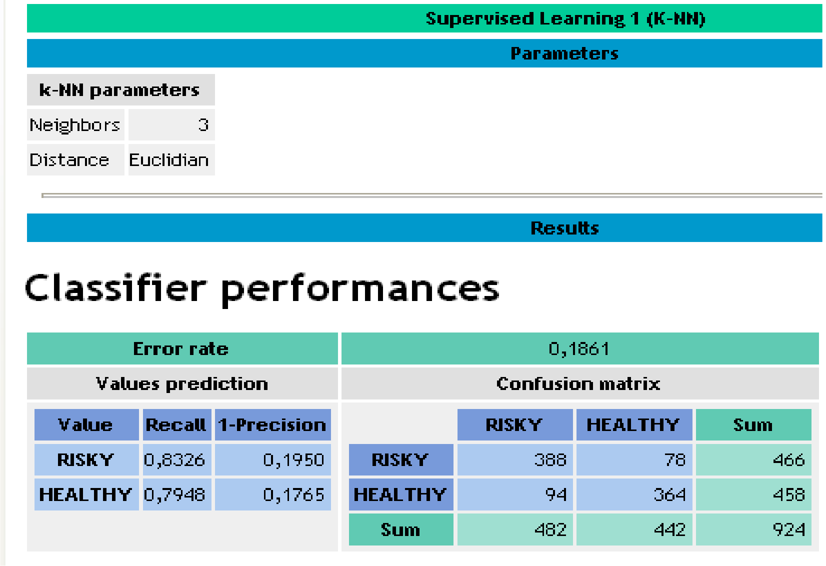
Values prediction

Value	Recall	1-Precision
RISKY	0,7854	0,2146
HEALTHY	0,7817	0,2183

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	366	100	466
HEALTHY	100	358	458
Sum	466	458	924

Panel 2: K-NN with k=3



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	3
Distance	Euclidian

Results

Classifier performances

Error rate	0,1861

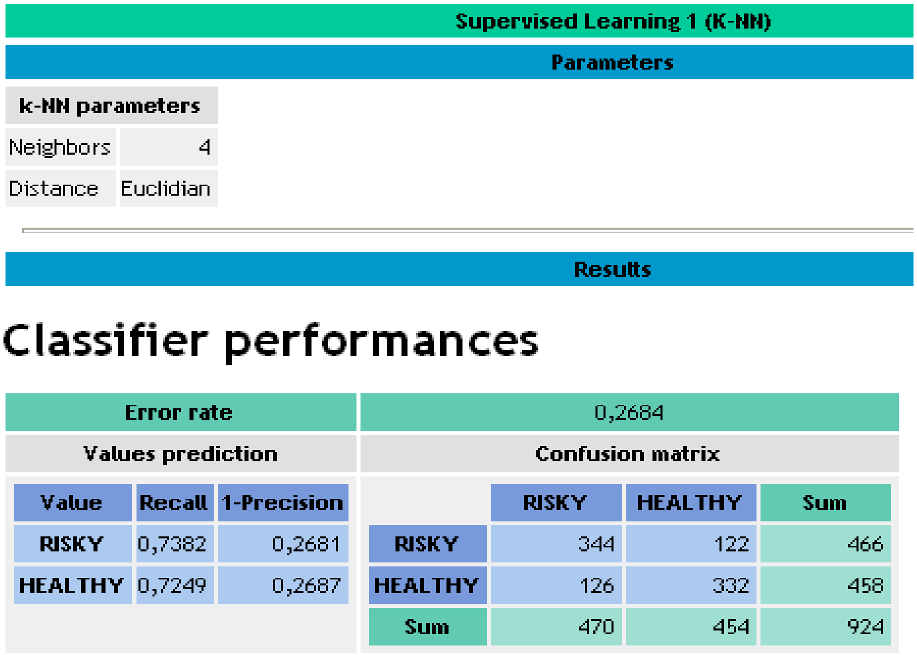
Values prediction

Value	Recall	1-Precision
RISKY	0,8326	0,1950
HEALTHY	0,7948	0,1765

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	388	78	466
HEALTHY	94	364	458
Sum	482	442	924

Panel 3: K-NN with k=4



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	4
Distance	Euclidian

Results

Classifier performances

Error rate	0,2684

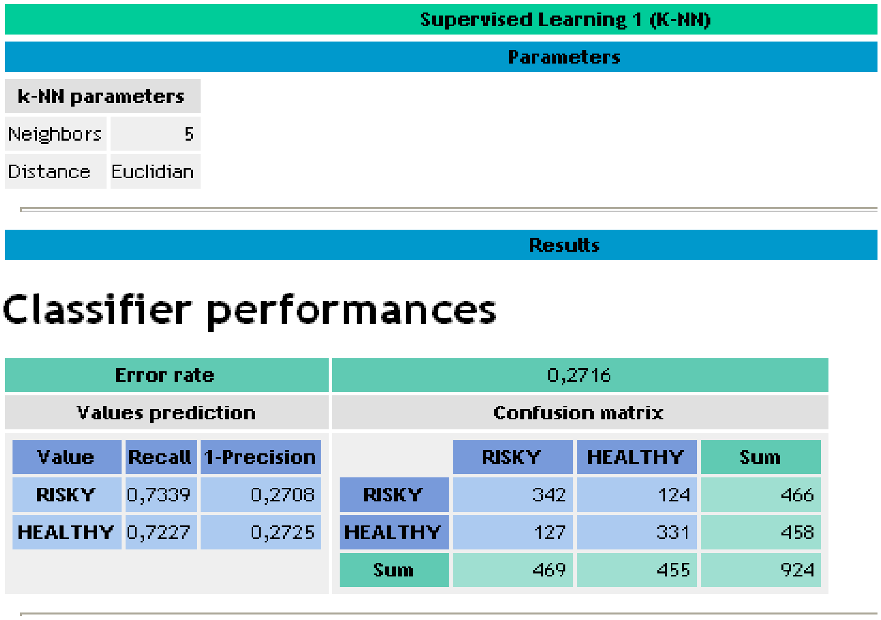
Values prediction

Value	Recall	1-Precision
RISKY	0,7382	0,2681
HEALTHY	0,7249	0,2687

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	344	122	466
HEALTHY	126	332	458
Sum	470	454	924

Panel 4: K-NN with k=5



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	5
Distance	Euclidian

Results

Classifier performances

Error rate	0,2716

Values prediction

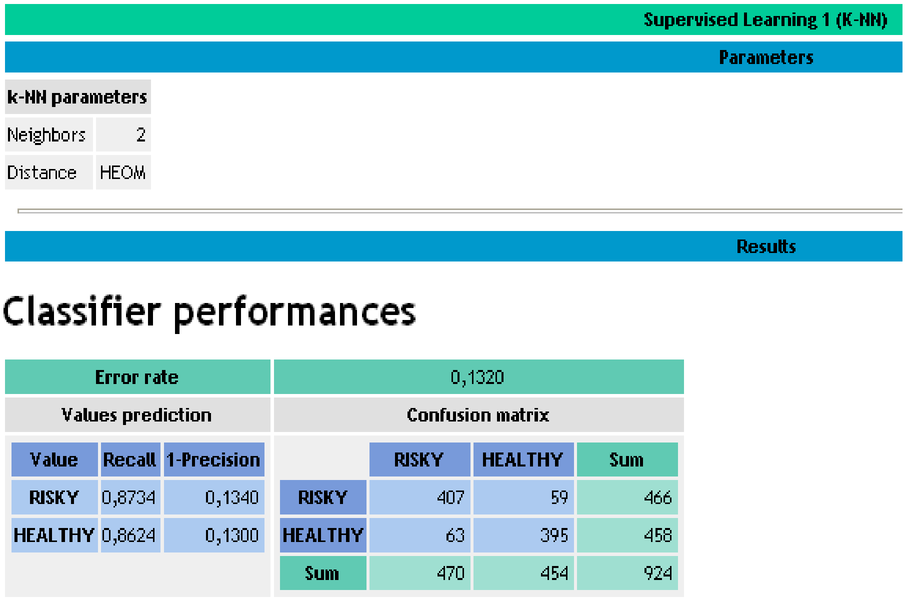
Value	Recall	1-Precision
RISKY	0,7339	0,2708
HEALTHY	0,7227	0,2725

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	342	124	466
HEALTHY	127	331	458
Sum	469	455	924

APPENDIX 2: CASH FLOW MODEL

Panel 1: K-NN with k=2



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	2
Distance	HEOM

Results

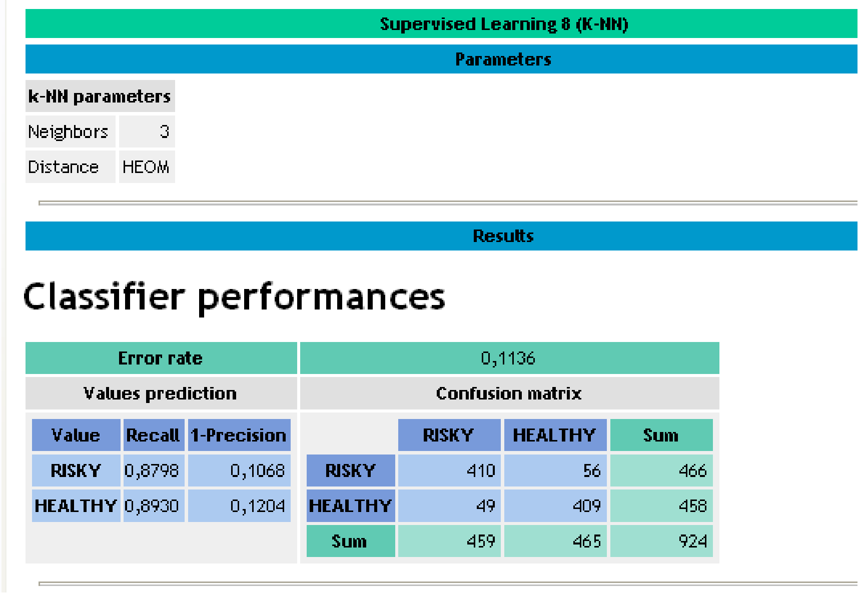
Classifier performances

Error rate	0,1320
Values prediction	Confusion matrix

Value	Recall	1-Precision
RISKY	0,8734	0,1340
HEALTHY	0,8624	0,1300

	RISKY	HEALTHY	Sum
RISKY	407	59	466
HEALTHY	63	395	458
Sum	470	454	924

Panel 2: K-NN with k=3



Supervised Learning 8 (K-NN)

Parameters

k-NN parameters	
Neighbors	3
Distance	HEOM

Results

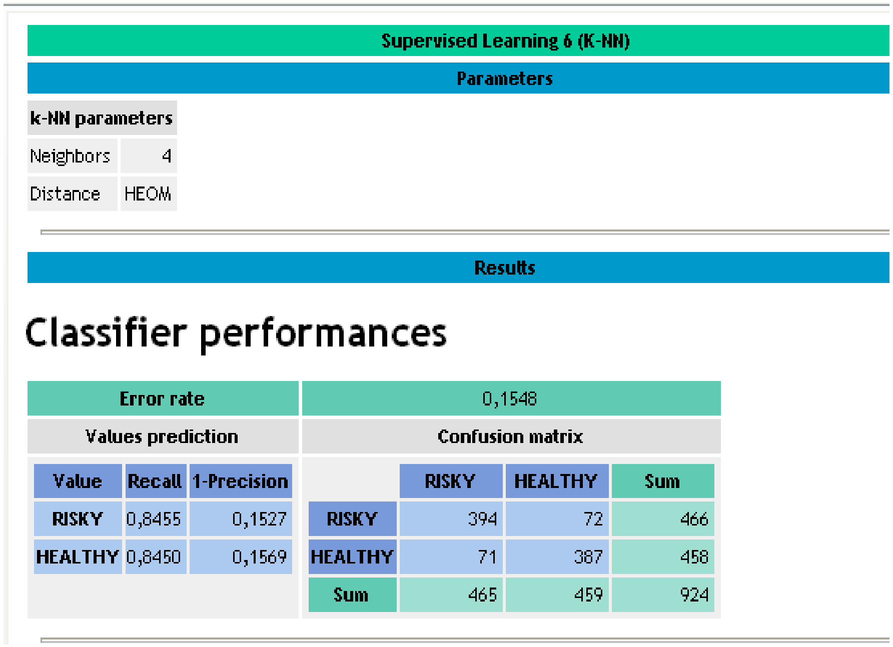
Classifier performances

Error rate	0,1136
Values prediction	Confusion matrix

Value	Recall	1-Precision
RISKY	0,8798	0,1068
HEALTHY	0,8930	0,1204

	RISKY	HEALTHY	Sum
RISKY	410	56	466
HEALTHY	49	409	458
Sum	459	465	924

Panel 3: K-NN with k=4



Supervised Learning 6 (K-NN)

Parameters

k-NN parameters	
Neighbors	4
Distance	HEOM

Results

Classifier performances

Error rate	0,1548

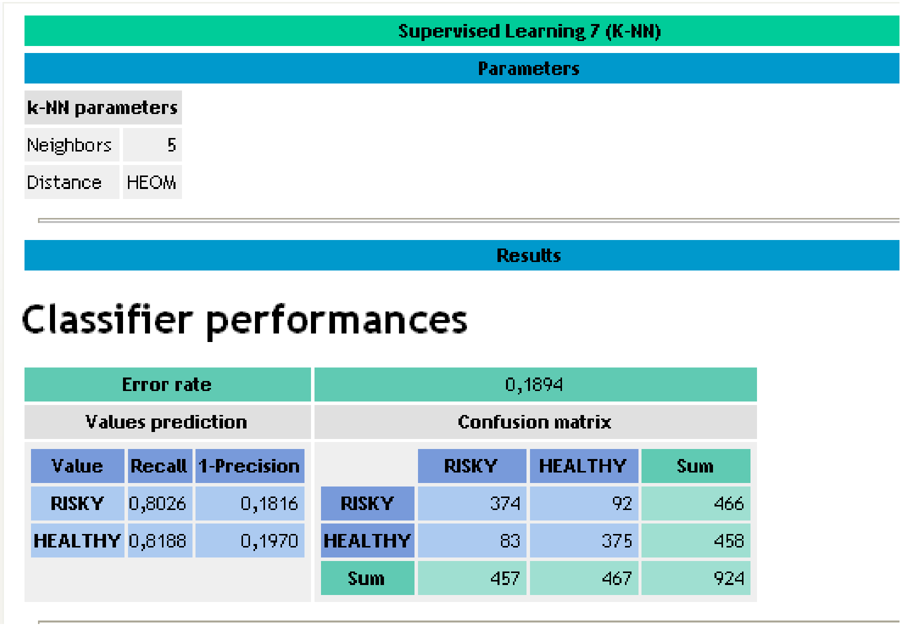
Values prediction

Value	Recall	1-Precision
RISKY	0,8455	0,1527
HEALTHY	0,8450	0,1569

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	394	72	466
HEALTHY	71	387	458
Sum	465	459	924

Panel 4: K-NN with k=5



Supervised Learning 7 (K-NN)

Parameters

k-NN parameters	
Neighbors	5
Distance	HEOM

Results

Classifier performances

Error rate	0,1894

Values prediction

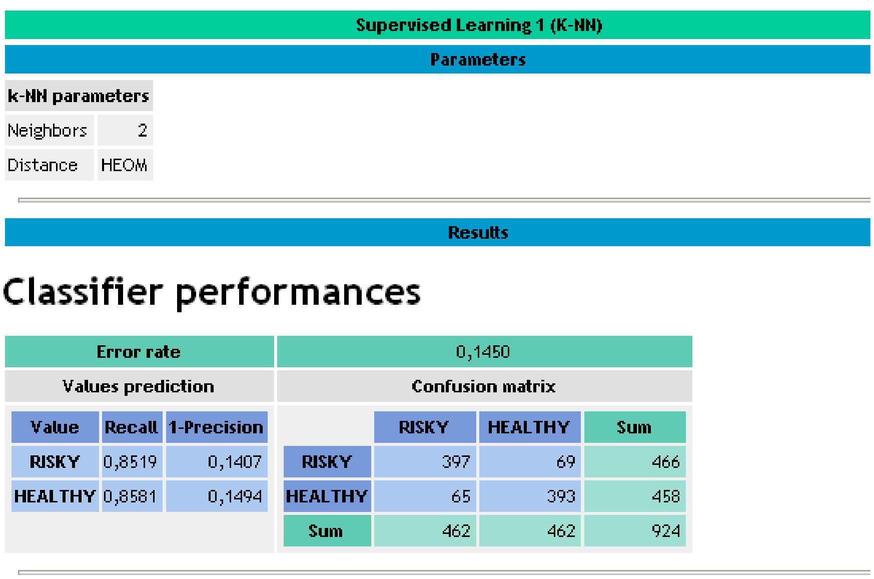
Value	Recall	1-Precision
RISKY	0,8026	0,1816
HEALTHY	0,8188	0,1970

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	374	92	466
HEALTHY	83	375	458
Sum	457	467	924

APPENDIX 3: FULL INFORMATION MODEL

Panel 1: K-NN with k=2



Supervised Learning 1 (K-NN)

Parameters

k-NN parameters	
Neighbors	2
Distance	HEOM

Results

Classifier performances

Error rate	0,1450

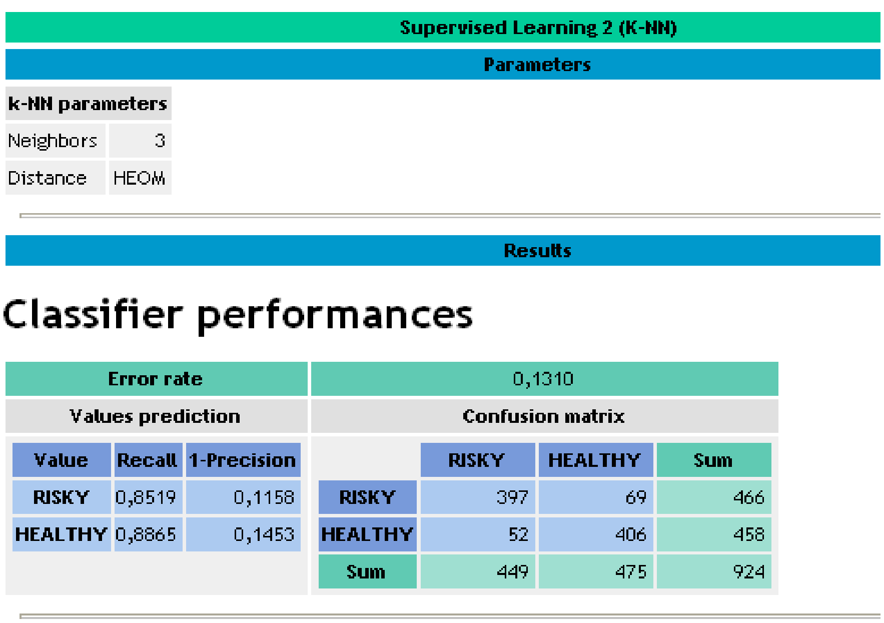
Values prediction

Value	Recall	1-Precision
RISKY	0,8519	0,1407
HEALTHY	0,8581	0,1494

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	397	69	466
HEALTHY	65	393	458
Sum	462	462	924

Panel 2: K-NN with k=3



Supervised Learning 2 (K-NN)

Parameters

k-NN parameters	
Neighbors	3
Distance	HEOM

Results

Classifier performances

Error rate	0,1310

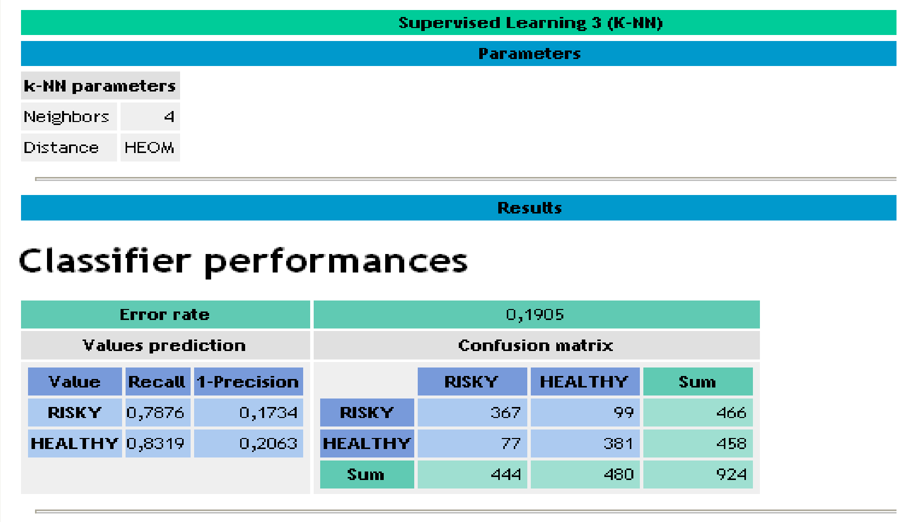
Values prediction

Value	Recall	1-Precision
RISKY	0,8519	0,1158
HEALTHY	0,8865	0,1453

Confusion matrix

	RISKY	HEALTHY	Sum
RISKY	397	69	466
HEALTHY	52	406	458
Sum	449	475	924

Panel 3: K-NN with k=4



Supervised Learning 3 (K-NN)

Parameters

k-NN parameters	
Neighbors	4
Distance	HEOM

Results

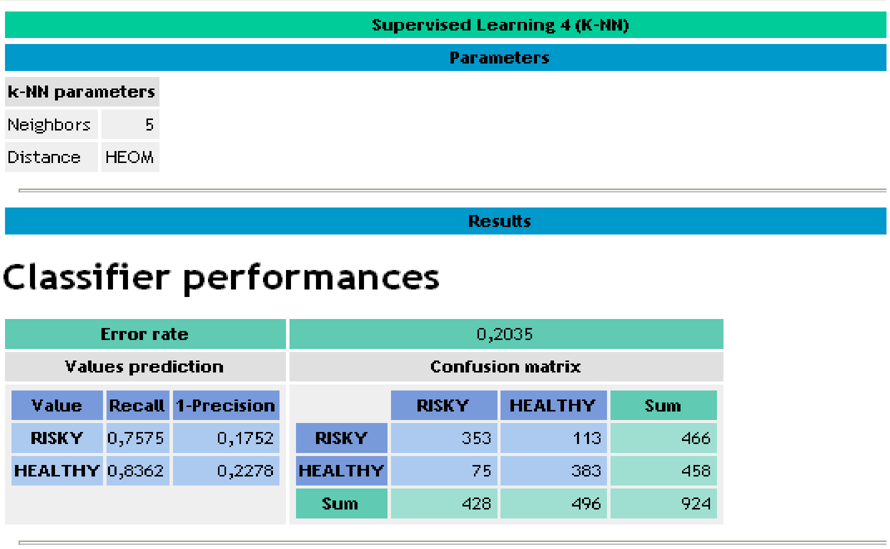
Classifier performances

Error rate	0,1905
Values prediction	Confusion matrix

Value	Recall	1-Precision
RISKY	0,7876	0,1734
HEALTHY	0,8319	0,2063

	RISKY	HEALTHY	Sum
RISKY	367	99	466
HEALTHY	77	381	458
Sum	444	480	924

Panel 4: K-NN with k=5



Supervised Learning 4 (K-NN)

Parameters

k-NN parameters	
Neighbors	5
Distance	HEOM

Results

Classifier performances

Error rate	0,2035
Values prediction	Confusion matrix

Value	Recall	1-Precision
RISKY	0,7575	0,1752
HEALTHY	0,8362	0,2278

	RISKY	HEALTHY	Sum
RISKY	353	113	466
HEALTHY	75	383	458
Sum	428	496	924

[1] "Sound Credit Risk Assessment and Valuation for Loans », Consultative Document, Bank for International Settlements Press & Communications, Basel (November 2005).

[2] *k*-Nearest Neighbors

[3] BIAT :Banque Internationale Arabe de Tunisie

[4] See (Karaa & Krichène, 2012 ; Matoussi & Abdelmoula, 2009)