The International Journal of Digital Accounting Research
Vol. 5, N. 9, 2005, pp. 1-45
ISSN: 1577-8517

Machine Learning and Statistical Techniques. An Application to the Prediction of Insolvency in Spanish Non-life Insurance Companies

Zuleyka Díaz. Universidad Complutense de Madrid. Spain.
zuleyka@ccee.ucm.es

María Jesús Segovia. Universidad Complutense de Madrid. Spain.
mjsegovia@ccee.ucm.es

José Fernández. Universidad Complutense de Madrid. Spain.
jfernan@ccee.ucm.es

Eva María del Pozo. Universidad Complutense de Madrid. Spain.
epozo@ccee.ucm.es

Abstract. Prediction of insurance companies insolvency has arisen as an important problem in the field of financial research. Most methods applied in the past to tackle this issue are traditional statistical techniques which use financial ratios as explicative variables. However, these variables often do not satisfy statistical assumptions, which complicates the application of the mentioned methods. In this paper, a comparative study of the performance of two non-parametric machine learning techniques (See5 and Rough Set) is carried out. We have applied the two methods to the problem of the prediction of insolvency of Spanish non-life insurance companies, upon the basis of a set of financial ratios. We also compare these methods with three classical and well-known techniques: one of them belonging to the field of Machine Learning (Multilayer Perceptron) and two statistical ones (Linear Discriminant Analysis and Logistic Regression). Results indicate a higher performance of the machine learning techniques. Furthermore, See5 and Rough Set provide easily understandable and interpretable decision models, which shows that these methods can be a useful tool to evaluate insolvency of insurance firms.

Key words: Insolvency, Insurance Companies, See5, Rough Set, Multilayer Perceptron, Discriminant Analysis, Logistic Regression.

Submitted June 2004
Accepted March 2005

1. INTRODUCTION

Insolvency, early detection of financial distress and conditions which lead to insolvency in insurance companies, are a concern for many insurance regulators, investors, management, financial analysts, banks, auditors, policy holders and consumers. It has been widely recognized that there should be some kind of supervision on such entities to attempt to minimize the risk of failure. Nowadays, *Solvency II* project is intended to lead to the reform of the existing solvency rules in the European Union.

Many insolvency cases appeared after the insurance cycles of the 1970s and 1980s in the United States and in the European Union. Several surveys have been devoted to identify the main causes of insurers' insolvency, in particular, the Müller Group Report (1997) analyses the main identified causes of insurance insolvencies in the European Union. These reasons can be summarized in: operational risks (operational failure related to inexperienced or incompetent management, fraud); underwriting risks (inadequate reinsurance programme and failure to recover from reinsurers, higher losses due to rapid growth, excessive operating costs, poor underwriting process) and insufficient provisions and imprudent investments.

The Solvency II project was initiated in 2001 to review the European framework for prudential supervision of insurance companies. It was divided in two phases. The first one tried to achieve a general design of the new supervisory regime. This phase ended at the beginning of 2003, and established that the new multi-pillar regulatory architecture adopted in the securities area was extended to insurance. The second phase of the project is currently under way. Undoubtedly, developing new methods to tackle prudential supervision in insurance companies is a highly topical issue for all countries that belong to the European Union, as in Spain's case.

A large number of methods have been proposed to predict business failure, however the special characteristics of the insurance sector have made most of them unfeasible, and just a few have been applied to this sector. Most approaches applied to prediction of failure in insurance companies are statistical methods such as Discriminant Analysis or Logistic Regression (Ambrose and Carroll, 1994; Bar-Niv and Smith, 1987; Mora, 1994; Sanchís *et al.*, 2003), which use financial ratios as explicative variables. In most cases this kind of variable does not usually

satisfy statistical assumptions. In order to avoid these problems, a number of non-parametric techniques have been developed, most of them belonging to the field of Machine Learning, such as neural networks (Serrano and Martín, 1993; Tam, 1991), which have been successfully applied to solve this kind of problems. However, their black-box character make them difficult to interpret, and hence the obtained results cannot be clearly analysed and related to the economical variables for discussion.

Other machine learning methods such as the ones tested in this paper (See5 and Rough Set) are more useful in economic analysis, because the models they provide can be easily understood and interpreted by human analysts. The purpose of this paper is to compare the predictive accuracy of these data analysis methodologies on a sample of Spanish non-life insurance companies, using general financial ratios and those that are specifically proposed for evaluating insolvency in insurance sector. Furthermore, in order to assess the efficiency of these methods, we will compare them with other widely used ones: Linear Discriminant Analysis, Logistic Regression and Multilayer Perceptron. The majority of previous researches have focused on the comparison of a certain method with the traditional statistical approaches (Altman *et al.*, 1994; De Andrés, 2001; Dimitras *et al*., 1999; Martínez de Lejarza, 1999), and only in few cases two or more machine learning techniques are compared with each other (Dizdarevic *et al.*, 1999; McKee and Lensberg, 2002; Salcedo *et al*., 2005).

The rest of the paper is structured as follows: in section 2, previous work on the prediction of insolvency in Spanish insurance sector is briefly reviewed. Section 3 introduces some concepts related to the tested techniques. In section 4 we describe the data and input variables. In section 5 the results of the five approaches are presented. The discussion and comparison of these results are also provided in this section. Section 6 summarizes some research limitations and, finally, section 7 closes the paper with some concluding remarks.

2. PREVIOUS WORK

Table 1 summarizes previous research that deals with the prediction of insolvency of Spanish insurance companies using financial ratios as explicative variables, and employing statistical techniques or artificial intelligence methods.

Author(s)	Technique(s)	Summary
López *et al.* (1994)	Discriminant Analysis	In this research the data sample consisted of 70 (35 failed and 35 non-failed) insurance firms in the 80's. Data five years prior to failure were collected. As a control measure, a failed firm is matched with a non-failed one in terms of line of business, business turnover and total assets. A linear discriminant function was developed for every year. The classification accuracies in percent of correctly classified firms for the five discriminant functions were: 90.85% (year 1), 76.56% (year 2), 74.60% (year 3), 70.97% (year 4) y 64.62% (year 5). The linear discriminant function developed for year 1 was used on the testing sample that consisted of 20 firms (10 failed and 10 non-failed). The classification accuracy in percent of correctly classified firms by this function on the testing sample was 80%. Yet the rest of functions have not been tested.
Martín *et al.* (1999)	Factor Analysis Cluster Analysis Discriminant Analysis	Multivariate analysis is applied to Spanish insurance sector for the period 1991-1994. The main objective was to develop a solvency ranking classification. In order to develop this classification, first of all a factorial analysis combined with a cluster analysis was carried out. This way, three groups that match with three solvency levels were obtained. Next step was to forecast the membership to one of the solvency levels by means of a linear discriminant function. For non-life insurance firms the classification accuracy in percent of correctly classified firms by the best discriminant function derived was 86.02%. For the mixed group the results were not very satisfactory (57%). In this research an external validation was not carried out, what could question the results reached by the discriminant functions developed.
Martínez de Lejarza (1999)	Multilayer Perceptron Discriminant Analysis	This research used the data from López *et al.* (1994) and a different forecasting model was developed for each year. A multilayer perceptron with two neurons in the hidden layer is trained for the five years. The classification accuracy in percent of correctly classified firms for the five years are: 100% for year 1, 97.96% for year 2, 96% for year 3, 100% for year 4 and 97.43% for year 5. Then a discriminant analysis is developed obtaining better results than the ones reached by López *et al.* (1994) though these results are worse than the ones obtained by the multilayer Perceptron: 92.31% of correctly classifications for year 1, 72.58% for year 2, 83.05% for year 3, 78.33% for year 4 and 73.68% for year 5. In this research an external validation is not carried out, so the results reached by both methods could be questionable.
Mora (1994)	Logistic Regression	A sample of 58 Spanish insurance firms (26 failed and 32 non-failed) is used. Three forecasting models for years 1, 2 and 3 before the firms went bankrupt were developed using logistic regression. The firms were classified into three categories: healthy firms, failed firms and uncertain firms. Considering these categories, the percentage of correctly classified firms for the training sample (20 healthy and 20 failed firms) was 95% (excluding the uncertain firms) for the three years. The percentages of correctly classified firms for the testing sample (12 healthy and 6 failed firms) were 83.33% for year 1, 77.78% for year 2 and 72.22% for year 3. None of the firms in the test sample is classified as uncertain firm.

Table 1. Previous Research

Sanchís *et al.* (2003)	Discriminant Analysis	Discriminant analysis is applied to a data sample consisted of 72 non-life insurance firms (36 failed and 36 non-failed). The firms were matched in terms of size (premiums volume). Taking as a starting point a set of 32 financial ratios, a *stepwise* procedure is used to perform a feature selection in the financial ratios space. In this research two types of discriminant functions (linear functions and quadratic functions) are developed using data five years prior to failure. The quadratic ones were developed due to covariance matrices were not equal, so results obtained by the linear models could be questioned. The forecasting results obtained by quadratic models were not satisfactory enough so the authors only considered the linear models. The percentages of correctly classified firms for the five linear discriminant functions were: 89.86% for year 1, 87.91% for year 2, 90.26% for year 3, 85.07% for year 4 y 94.44% for year 5. A cross-validation procedure is used to validate the results. The percentages of correctly classified firms were: 81.86% for year 1, 81.27% for year 2, 76.79% for year 3, 75.34% for year 4 and 77.78% for year 5.
Segovia *et al.* (2004)	Support Vector Machines -SVM Genetic Algorithms - GA Simulated Annealing - SA	This research used the sample from Sanchís *et al.* (2003) but other financial ratios were calculated. A SVM is used to classify firms and both GA and SA are used to perform on-line feature selection in the ratios space. The percentage of correctly classified firms for the first year before failure using the SVM without feature selection is 67% (using cross-validation procedure). The feature selection using GA and SA provided two sets containing only three ratios instead of 19 initial ones. The percentage of correctly classified firms using the SVM with feature selection (for the two sets) is 77% (using cross-validation procedure).

Table 1 (Continued). Previous Research

3. A BRIEF OVERVIEW OF THE TESTED TECHNIQUES

3.1. The See5 algorithm

Learning systems based on decision trees are known to be the easiest to use and understand among all machine learning methods. Moreover, the condition and ramification structure of a decision tree is suitable for classification problems. Prediction of insolvency is a kind of classification problem, as we try to classify firms into solvent or insolvent.

The automatic construction of decision trees begins with the studies developed in the social sciences by Morgan and Sonquist (1963) and Morgan and Messenger (1973). In statistics, the CART (Classification and Regression Trees) algorithm to generate decision trees proposed by Breiman et al. (1984) is one of the most important contributions. At around the same time decision tree induction was beginning to be used in the field of machine learning, notably by Quinlan (1979, 1983, 1986, 1988, 1993 and 1997), and in engineering by Henrichon and Fu (1969) and Sethi and Sarvarayudu (1982).

The successive branches of a decision tree achieve a series of exhaustive and mutually exclusive partitions among the set of objects that a decision maker wants to classify. The main difference among the different algorithms used is the criterion followed to carry out the partitions previously mentioned.

The See5 algorithm (Quinlan, 1997) is the latest version of the ID3 and C4.5 algorithms developed by this author in the last two decades. The criterion employed in See5 algorithm to carry out the partitions is based on some concepts from Information Theory and has been improved significantly over time. The main idea shared with similar algorithms is to choose the variable that provides more information to realize the appropriate partition in each branch, in order to classify the training set.

The information provided by a message or a random variable x is inversely proportional to its probability (Reza, 1994). This quantity is usually measured in *bits* obtained through the relation: $\log_2 \frac{1}{p_x}$. The average of this relation for all the possible cases of the random variable x is called *entropy* of x: $H(x) = \sum_x p(x) \log_2 \frac{1}{p(x)}$. The entropy is a measure of the randomness or uncertainty of x or a measure of the average amount of information that is supplied by the knowledge of x.

In the same way, we can define the *joint entropy* of two random variables x and y: $H(x,y) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x,y)}$, which represents the average amount of information supplied by the knowledge of x and y. The *conditional entropy* of x given the variable y, $H(x/y)$, is defined as $H(x/y) = \sum_{x,y} p(x,y) \log_2 \frac{1}{p(x/y)}$, and this relation is a measure of the uncertainty of x when we know the variable y. This is the amount of information necessary to know completely x when we know the information provided by y-variable. Naturally, $H(x/y) \leq H(x)$,

because if *y*-variable is known we have more information that can help us to reduce the uncertainty about *x*-variable. This reduction in the uncertainty is called *mutual information* between *x* and *y*: $I(x;y) = H(x) - H(x/y)$, which is the information provided by one of the variables about the other one. It is always verified that $I(x;y) = I(y;x)$, therefore the amount of information that each variable provides about the other one is the same.

We can consider that *x* is a random variable that represents the category to which an object belongs. On the other hand, y_i, $i = 1, 2, ..., n$, represents the set of attributes that describe the objects we want to classify.

See5 algorithm chooses to make each partition the y_i-variable that provides the maximum information about *x*-variable, that is, it maximizes the following relation called *gain ratio*: $\frac{I(x;y_i)}{H(y_i)}$. This ratio represents the percentage of information provided by y_i that is useful in order to characterize *x*.

Note that $I(x;y_i)$ should be large enough to prevent that an attribute could be only chosen because it has a low value for entropy, what would increase the *gain ratio*.

A common problem for most of rules and tree induction systems is that models they generate can be quite adapted to the training set, so the classification obtained will be nearly perfect. Consequently, the model developed will be very specific and if we want to classify new objects, the model will not provide good results, especially if the training set has noise. In this last case, the model would be influenced by errors (noise) which would lead to a lack of generalization. This problem is known as *overfitting*.

The most frequent way of limiting this problem in the context of decision trees consists in deleting some conditions of the tree branches, in order to achieve more general models. This procedure can be considered as a *pruning* process. This way we will increase the misclassifications in the training set but, at the same time, we probably decrease the misclassifications in the test set that has not been used to develop the decision tree.

Quinlan incorporates a *post-pruning* method for an original fitted tree. This method consists in replacing a branch of the tree by a leaf, conditional on a predicted error rate. Suppose that there is a leaf that covers *N* objects and misclassifies *E* of them. This could be considered as a binomial distribution in which the experiment is repeated *N* times obtaining *E* errors. From this issue, the probability of error p_e is estimated, and it will be taken as the aforementioned predicted error rate. So it is necessary to estimate a confidence interval for the error probability of the binomial distribution. The upper limit of this interval will be p_e (note that this is a pessimistic estimate).

Then, in the case of a leaf that covers *N* objects, the number of predicted errors will be $N \cdot P_e$. If we consider a branch instead of a leaf, the number of predicted errors associated with a branch will be just the sum of the predicted errors for its leaves. Therefore, a branch will be replaced by a leaf when the number of predicted errors for the last one is lower than the one for the branch.

Furthermore, See5 algorithm includes additional functions such as a method to change the obtained tree into a set of classification rules that are generally easier to understand than the tree. For a more detailed description of the features and workings of See5 algorithm see Quinlan (1993 and 1997).

3.2. Rough set theory: main concepts

Rough Set Theory (RS Theory) was firstly developed by Pawlak (1991) in the 1980s as a mathematical tool to deal with the uncertainty or vagueness inherent in a decision making process. Though nowadays this theory has been extended (Greco *et al*, 1998), we refer to the classical approach. RS Theory is somewhat different to probability theory, which deals with random events in nature or fuzzy set theory, which deals with objects that may belong to more than one category in different degrees.

On the other hand, RS Theory is very well fitted when the classes into which the objects have to be classified are imprecise but can be approximate with precise sets (Nurmi *et al.*, 1996). Therefore, these differences show one of the main advantages of this theory: an agent is not required to assign precise numerical values to express imprecision of his knowledge, such as probability distributions in statistics or grade of membership in fuzzy set theory (Pawlak, 1991).

This section presents some concepts of RS Theory following Pawlak´s reference and some remarks by Slowinski (1993) and Dimitras *et al.* (1999).

This approach is based on the assumption that with every object of the universe we are considering we can associate knowledge, data. Knowledge is regarded as ability to classify objects. Objects described by the same data or knowledge are indiscernible in view of such knowledge. The *indiscernibility* relation leads to mathematical basis for the RS Theory. Intuitively, a rough set is a set or a subset of objects that cannot be expressed exactly by employing available knowledge. If this information or knowledge consists of a set of objects described by another set of attributes, we consider a rough set as a collection of objects that cannot be precisely characterized in terms of the values of the set of attributes.

RS Theory represents knowledge about the objects as a data table. Rows are labelled by objects (states, processes, firms, patients, candidates,…) and columns are labelled by attributes. Entries of the table are attribute values. Therefore, for each pair object-attribute, x-q, there is known a value called *descriptor*, $f(x, q)$. The *indiscernibility relation* would occur if for two objects, x and y, all their descriptors in the table have the same values, that is, if and only if $f(x, q) = f(y, q)$.

3.2.1. Accuracy and quality of approximation

Any rough set has a lower and an upper approximation in terms of classes of indiscernible objects. Thus, a rough set is a collection of objects that, in general, cannot be precisely characterized in terms of the values of the set of attributes, while its lower and upper approximations can. The lower approximation consists of all objects which certainly belong to the set and can be certainly classified as elements of that set, using the set of attributes in the table (the knowledge we are considering). The upper approximation contains objects which possibly belong

to the set and can be possibly classified as elements of that set using the set of attributes in the table. The *boundary* or *doubtful region* is the difference between the lower and the upper approximation and this is the set of elements which cannot be certainly classified using the set of attributes.

The quotient between the cardinality of the lower approximation and the cardinality of the upper one represents the percentage of possible correct decisions when classifying objects using knowledge available.

As we are interested in classifying a set of objects, the *quality of classification* is defined as the quotient between the addition of the cardinalities of the lower approximations of all the classes in which the objects are classified, and the number of these objects.

3.2.2. Reduction and dependency of attributes

A fundamental problem in the rough set approach is discovering dependencies between attributes in an information table, because it allows to reduce the set of attributes removing those that are not essential (unnecessary) to characterize knowledge. This problem will be referred to as knowledge reduction, and the main concepts related to this question are the *core* and the *reduct.* A reduct is the minimal subset of attributes which provides the same quality of classification as the set of all attributes. If the information table has more than one reduct, the intersection of all of them is called the core and is the collection of the most relevant attributes in the table.

3.2.3. Decision rules

An information table which contains condition and decision attributes is referred as a decision table. A decision table specifies what decisions (actions) should be undertaken when some conditions are satisfied. So a reduced information table may provide decision rules of the form "*if* conditions *then* decisions".

These rules can be *deterministic* when the rules describe the decisions to be made when some conditions are satisfied and *non-deterministic* when the decisions are not univocally determined by the conditions so they can lead to several possible

decisions if their conditions are satisfied. The number of objects that satisfy the condition part of the rule is called the *strength* of the rule and is a useful concept to assign objects to the strongest decision class when rules are non-deterministic.

The rules derived from a decision table do not usually need to be interpreted by an expert as they are easily understandable by the user or decision maker. The most important result in this approach is the generation of decision rules because they can be used to assign new objects to a decision class by matching the condition part of one of the decision rule to the description of the object. So rules can be used for decision support.

RS Theory can analyse several multiattribute decision problems. It is especially well suited to sorting problems. One of these problems is multiattribute sorting problem which consists in the assignment of each object, described by values of attributes, to a predefined class or category. Business failure is an example of this kind of problem as we try to assign firms (objects) described by a set of financial ratios (attributes) to a category ("failed" or "healthy" firm).

3.3. Multilayer perceptron

Within the framework of neural networks, the Multilayer Perceptron is one of the most widely used problem-solving architectures in a great variety of areas, thanks, largely, to its proficiency as an universal approximator of non-linear relationships between data input and output. In addition, it is easy to use and apply.

Multilayer Perceptron is an advance on simple Perceptron (Rosenblatt, 1957) and arose in response to some limitations found in the simple version of the architecture. In 1986, Rumelhart, Hinton and Williams (Rumelhart *et al.*, 1986) formalized a method through which a neuronal network could learn the existing association between the input patterns and the corresponding outputs, utilizing more levels of neurons than Rosenblatt used to develop the Perceptron. This method, known as *backpropagation* (backward error propagation), is an extension to networks with intermediate layers (multilayer networks) and non-linear activation functions of the Delta rule proposed by Widrow and Hoff (1960) to account for the error produced by exits from the network.

The importance of the backpropagation network stems from the internal representation of the knowledge that can be organized in the intermediate layer of cells for the purpose of accomplishing any correspondence between input and output in the network, self-adapting the weights of the neurons in the intermediate layers.

Very briefly, the workings of the backpropagation network consists in learning from a set of input-output pairs by means of the following process: first, an input pattern is applied as a stimulus for the first layer of neurons of the network, which continues propagating through all the adjacent layers until generating an output, and the results obtained in the output neurons are compared with the desired output and an error value is calculated for each output neuron. Next, these errors are transmitted backwards, starting from the exit layer, toward all the neurons of the intermediate layer that contribute directly to the output, receiving the percentage of error that corresponds to the participation of the intermediate neuron in the original output. This process continues, layer by layer, until all the neurons of the network have received an error that describes their relative contribution to the total error. Based on the value of the error received, the weights of the connections between the neurons are readjusted. Thus, the next time the same pattern occurs the output will be closer to the desired value and in this way the error decreases. In successive cycles the parameters of the network are adjusted until the error reaches a minimum.

The ability of the Multilayer Perceptron to approximate non-linear functions, to filter noise in the data, etc., makes it an appropriate model to handle real problems. Nevertheless, while it is one of the most well-known and used networks, this does not imply that it is one of the most potent or that it offers the best results in different areas of application.

In spite of the great predictive efficiency that neural networks have shown in numerous empirical studies, we should mention that they are “black box” models and involve serious difficulties of theoretical interpretation. Thus, their utilization would only be advisable in those situations where explanation is less important than prediction.

3.4. Linear discriminant analysis (LDA)

LDA is one of the best-known and most utilized classification techniques. It consists of a series of linear functions of observations, called discriminant functions, which allow dividing the space of the classification variables in a group of regions separated by linear boundaries. The region in which each observation falls determines the class to which it is assigned. In our case, having two different classes, the space will be divided into two regions separated by a hyperplane, one corresponding to healthy firms and the other to failed firms.

LDA is an optimal classification method in the sense that it minimizes the probability of an erroneous classification of new observations. To do this however requires certain restrictive hypotheses to be fulfilled. Namely, the classification variables should follow a normal multivariant distribution and the covariance matrixes for the observations of each class should be equal (homoscedasticity). If these requirements are not met, LDA is not the best possible classifier, but it can still be used and offers good results in many cases. This is because LDA can be considered a suitable method to search for projection directions that maximize the separation between elements of different classes and this purely geometric interpretation is not affected by hypotheses on the distribution of data.

In actual practice, if the hypotheses on normality and homoscedasticity are not fulfilled, it is not easy to determine beforehand whether LDA or an alternative technique, like Logistic Regression, will provide better results. Therefore, the best answer to the problem is usually to compare the results afterwards (Peña, 2002; Webb, 2002).

3.5. Logistic regression

If the hypotheses on normality and homoscedasticity that would allow LDA to provide optimal results are not fulfilled, it could be wise to use Logistic Regression as the classification method. Although it does not always surpass the usefulness of LDA, Logistic Regression is usually more efficient when the populations have different covariance matrixes or are distinctly non-normal.

Logistic Regression consists of making a Maximum Likelihood Estimation of the parameters of a linear function of the explicative variables. That linear function

provides estimations of the magnitude $\log\frac{p}{1-p}$, where p will be the probability of a random binary variable that follows a Bernoulli distribution. The values that this variable takes indicate the class which each observation belongs to. Given a new observation characterized by certain concrete values from $x_1, x_2, ..., x_p$ the model gives us the estimated probability of this observation belonging to one class or another.

4. METHODOLOGICAL ASPECTS

4.1. Selection of data and variables

In this section, we show the main characteristics of the data and variables that will be used to develop our models. We have used the sample of Spanish firms used by Sanchís *et al.* (2003) in the application of the Discriminant Analysis for the prediction of failure in non-life insurance companies. This data sample consists of non-life insurance firms data five years before failure. The firms were in operation or went bankrupt between 1983 and 1994. From this period, 72 firms (36 failed and 36 non-failed) are selected. As a control measure, a failed firm is matched with a non failed one in terms of industry and size (premiums volume), following the methodology developed by other authors in similar applications of the Discriminant Analysis: Altman (1968); Altman *et al.* (1977); López *at al.* (1994); Martínez de Lejarza (1999); Mora (1994). Furthermore, the firm size is a so important variable for the prediction of insolvency that its inclusion could cloud the role of other financial variables which we are especially interested in.

We have developed three models using data of one, two and three years before the firms declared bankruptcy. Thus, it has to be noted that the prediction of the insolvency achieved by each of them will be one, two and three years in advance, respectively. We refer to these models as *Model 1*, *Model 2* and *Model 3*.

In order to test the predictive accuracy of the models, we have split the set of original data to form the training sets and the holdout samples to validate the obtained models, i.e., the test sets. For *Model 1*, the training set consisted of 54 firms (27 failed and 27 non-failed firms) randomly generated. Therefore we have

left 18 firms (9 failed and 9 non-failed) for testing. Sample size is different each year from the others, because data didn't exist for all the firms. In Table 2 these sample sizes are shown as well as the sizes of the training sets (randomly generated) to develop the models and the test sets to validate them.

Model	Sample size (number of firms)	Training set (number of firms)	Test set (number of firms)
1	72 (36 failed and 36 non-failed)	54 (27 failed and 27 non-failed)	18 (9 failed and 9 non-failed)
2	68 (34 failed and 34 non-failed)	52 (26 failed and 26 non-failed)	16 (8 failed and 8 non-failed)
3	54 (27 failed and 27 non-failed)	40 (20 failed and 20 non-failed)	14 (7 failed and 7 non-failed)

Table 2. Sample Sizes

Each firm is described by 21 financial ratios that have come from a detailed analysis of the variables and previous bankruptcy studies for insurance companies. Appendix A shows the 21 ratios which describe the firms. Note that special financial characteristics of insurance companies require general financial ratios as well as those that are specifically proposed for evaluating insolvency of insurance sector.

The ratios have been calculated from the financial statements (balance sheets and income statements) issued one, two and three years before the firms were declared bankrupt. Ratios 15 and 16 have been removed in our study due to the fact that most of the firms do not have "other income" so there is no sense in using them for an economic analysis. This reduces the total number of ratios to 19.

4.2. Implementation of the proposed techniques

Linear Discriminant Analysis and Logistic Regression have been performed using *R 2.0.0* software distributed by CRAN Foundation (R Development Core Team, 2004). The software used to implement See5 algorithm is *See5* by RULEQUEST RESEARCH (Quinlan, 1997). The Multilayer Perceptron has been performed using the data mining package *WEKA* from the University of Waikato (Witten and Frank, 2000). And, finally, Rough Set analysis has been performed using *ROSE* software provided by the Institute of Computing Science of Pozna University of Technology (Predki *et al.* (1998); Predki and Wilk (1999)).

5. RESULTS

5.1. See5 algorithm

We have developed three models (three decision trees). We refer to them as *Model 1*, *Model 2* and *Model 3*. They have been developed using, respectively, the previously mentioned training sets 1, 2 and 3, and we have tested them with the test sets 1, 2 and 3. Next, *Model 1* is shown:

Model 1

```
R13 > 0.68:
:...R9 <= 0.59: failed (14)
:   R9 > 0.59:
:   :...R17 <= 0.99: failed (3)
:       R17 > 0.99: healthy (3)
R13 <= 0.68:
:...R1 > 0.29: healthy (20/2)
    R1 <= 0.29:
    :...R2 > 0.04: failed (3)
        R2 <= 0.04:
        :...R6 > 0.64: healthy (3)
            R6 <= 0.64:
            :...R9 <= 0.85: failed (4)
                R9 > 0.85: healthy (4/1)



Evaluation on training data (54 cases):

     Decision Tree
   ----------------
   Size      Errors

      8     3(5.6%)

    (a)   (b)    <-classified as
   ----  ----
     27          (a): class healthy
      3    24    (b): class failed
```

```
Evaluation on test data (18 cases):

     Decision Tree
   ----------------
   Size  Errors
      8    5 (27.8%)
    (a)   (b)    <-classified as
   ----  ----
      7     2    (a): class healthy
      3     6    (b): class failed
```

As we can see, only 6 ratios appear in the tree instead of the 19 initial ones. This indicates that these 6 variables are the most relevant ones for discrimination between solvent and insolvent firms in our sample and, consequently, it shows the strong support of this approach in feature selection. Our tree would be read in the following way:

- If the ratio R13 is greater than 0.68 and the ratio R9 is less than or equal to 0.59, then the company will be classified as "failed". This fact is verified by 14 firms in our sample.

- If the ratio R13 is greater than 0.68 and the ratio R9 is greater than 0.59 and the ratio R17 is less than or equal to 0.99, then the company will be classified as "failed", completing these conditions 3 companies.

- If...

and so on.

Every leaf of the tree is followed by a number n or n/m. The value of n is the number of cases in the sample that are mapped to this leaf, and m (if it appears) is the number of them that are classified incorrectly by the leaf.

The section under the tree concerns the evaluation of the decision tree, first on the cases of the training set from which it was constructed, and then on the new cases of the test set. The size of the tree is its number of leaves and the column headed "Errors" shows the number and percentage of cases misclassified. The tree, with 8 leaves, misclassifies 3 of the 54 given cases, what implies an error rate of 5.6%, that is, 94.4% of correctly classified firms. Performance on the training cases is further analyzed in a *confusion matrix* that pinpoints the kinds of errors

made. A similar report of performance is given for the test cases, that shows the model's accuracy on unseen test cases: an error rate of 27.8%, that is, 72.2% of correctly classified firms.

Though the tree we have developed is quite easy to understand, sometimes the trees developed are difficult to interpret. An important feature of See5 is its ability to generate unordered collections of *if-then* rules, which are simpler and easier to understand than decision trees. The rules that are obtained starting from the previous tree are:

```
Rule 1: (20/2, lift 1.7)
 R1 > 0.29
 R13 <= 0.68
 ->  class healthy  [0.864]

Rule 2: (12/1, lift 1.7)
 R2 <= 0.04
 R6 > 0.64
 R13 <= 0.68
 ->  class healthy  [0.857]

Rule 3: (7/1, lift 1.6)
 R9 > 0.85
 ->  class healthy  [0.778]

Rule 4: (14, lift 1.9)
 R9 <= 0.59
 R13 > 0.68
 ->  class failed  [0.938]

Rule 5: (7, lift 1.8)
 R13 > 0.68
 R17 <= 0.99
 ->  class failed  [0.889]

Rule 6: (26/6, lift 1.5)
 R1 <= 0.29
 ->  class failed  [0.750]


Default class: healthy
```

Each rule consists of:

- Statistics (*n*, lift *x* or *n/m* lift *x*) that summarize the performance of the rule. Similarly to a leaf, *n* is the number of training cases covered by the rule and *m*, if it appears, shows how many of them do not belong to the class predicted by the rule. The lift *x* is the result of dividing the estimated accuracy of the rule by the relative frequency of the predicted class in the training set. The accuracy of the rule is estimated by the Laplace ratio (*n*-*m*+1)/(*n*+2) (Clark and Boswell, 1991; Niblett, 1987).

- One or more conditions that must all be satisfied if the rule is to be applicable.

- A class predicted by the rule.

- A value between 0 and 1 that indicates the confidence with which this prediction is made.

There is also a *default class*, here "healthy", which is used when an object does not match any rule.

In this model, performance on the training cases and on the test cases is the same with this ruleset as with the previous tree, but it won't always be this way.

Although these results are satisfactory, they can be improved by applying the *boosting* option that See5 incorporates, based on Freund and Schapire's research (1997). Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy. Very briefly, the idea is to create several classifiers (either decision trees or rulesets) rather than just one. As the first step, a single decision tree or ruleset is built as before from the training data. This classifier will usually make mistakes on some cases in the data. When the second classifier is built, more attention is paid to these cases in an attempt to get them right. As a consequence, the second classifier will generally be different from the first. It will also make errors on some cases, and these will become the focus of attention during the construction of the third classifier. This process continues for a pre-determined number of iterations or *trials*. Finally, when a new case is to be classified, each classifier votes for its predicted class and the votes are counted to determine the

final class. The results obtained with this method are frequently very good.

In this way, starting from the previous tree, the results reached by means of the boosting option with 18 trials are shown in Table 3, in percent of correctly classified firms.

Correct classifications	Training set	Test set
"healthy" firms	100%	77.78%
"failed" firms	100%	88.89%
Total	100%	83.33%

Table 3. Boosting Results for *Model 1*

The sets of variables in the trees that constitute the rest of the models are shown in Table 4. This table also displays performance on the training cases and on the test cases, in percent of correctly classified firms. The trees 2 and 3 have been pruned, because previously we observed that the error rates were quite smaller on the training sets than on the test sets, and this could be due to an overfitting problem. However, pruning doesn't improve performance on the first tree.

Model	Set of variables	Size of the tree	Correct classifications			
			Training set		Test set	
			"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	R13, R9, R17, R1, R2, R6	8	100%	88.89%	77.78%	66.77%
			Total: 94.44%		Total: 72.22%	
2	R1, R13, R20, R7, R3	6	96.15%	84.62%	87.5%	75%
			Total: 90.39%		Total: 81.25%	
3	R4, R19, R1	5	100%	70%	100%	57.14%
			Total: 85%		Total: 78.57%	

Table 4. See5 Results

As we previously mentioned, in many occasions the classification accuracy can be improved by means of boosting. For example, for model 2, the results that we have obtained by means of boosting with 11 trials are shown in Table 5, in percent of correctly classified firms.

Correct classifications	Training set	Test set
"healthy" firms	100%	87.5%
"failed" firms	100%	87.5%
Total	100%	87.5%

Table 5. Boosting Results for *Model 2*

5.2. Rough set

The first analysis we have made is to recode the ratios (continuous variables) into qualitative terms (low, medium, high and very high) with corresponding numeric values such us 1, 2, 3 and 4. The recoding has been done dividing the original domain into subintervals. This recoding is not imposed by the RS theory but it is very useful in order to draw general conclusions from the ratios in terms of dependencies, reducts and decision rules (Dimitras *et al.*, 1999).

We have decided to recode the information tables using 4 subintervals based on the quartiles for the actual ratios values (years 1, 2 and 3) for the whole samples. The list of subintervals for the first year is shown in Appendix B.

We have used the subintervals assigning the highest code to the best subinterval to develop a coded information table, thus for the ratios for which lower values are better, we have given the codes in the inverse order of the subintervals. Moreover, RS Theory allows us to make corrections on the scale if our experience or knowledge is not concordant with the increasing or decreasing sequence of subintervals. For example, experience in insurance sector demonstrates that for ratios R5 to R10 the best percentiles correspond to the central part of the distribution, and it is preferable to be in the third percentile than in the second one. Therefore we have made corrections in the scale for these ratios. We have also made corrections for ratios R11 to R19. The assignment of codes to quartiles for the first year is presented in Appendix C.

The first results of the analysis indicated that the approximation of the decision classes and their quality of classification were equal to one and the core of attributes was empty. These results show that the firms are very well discriminated (so the boundary regions are empty for the two decision classes) and that none of the attributes is indispensable for the approximation of the two decision classes.

Next step of the Rough Set analysis was the generation of the reducts. For example, for *Model 1* we have obtained 229 reducts which contain 4-7 attributes.

These results mean that at least 12 attributes are redundant (and, therefore, they could be eliminated). We have selected the reducts consisted of R3, R4, R9, R14 and R17, for *Model 1*, R1, R3, R4, R5 and R17, for *Model 2*, and R3, R4, R14 and R17, for *Model 3*, taking into account three questions:

- The number of attributes should be as small as possible.
- It should have the most significant attributes in our opinion for the evaluation of the companies.
- After having selected a few reducts containing the most significant attributes, the reduct chosen should not contain ratios with a very high value of the correlation coefficients.

Once we have chosen a reduct, the rest of attributes of a coded information table can be eliminated. The reduced tables will be used to obtain the decision rules. The strategy we have followed to obtain the decision rules consists in the generation of a minimal subset of rules covering all the objects from the decision table (so the correct classifications on training sets will be always 100%). This strategy is implemented in the ROSE software.

We have obtained three algorithms: *Model 1* consists of 27 rules (see Appendix D), *Model 2* consists of 25 rules and *Model 3* consists of 22 decision rules. All of them are deterministic because the quality of the classification is equal to 1 and this means that the doubtful region is empty.

The models have been tested on data from the test sets, i.e., on the rest of firms that have not been used to estimate the algorithms. The classifications accuracies in percent of correctly classified firms are shown in Table 6.

Model	Set of variables (reduct)	Number of decision rules	Correct classifications	
			"Healthy" firms	"Failed" firms
1	R3, R4, R9, R14, R17	27	77.78%	77.78%
2	R1, R3, R4, R5, R17	25	Total: 77.78%	
			75%	75%
3	R3, R4, R14, R17	22	Total: 75%	
			71.43%	57.14%
			Total: 64.29%	

Table 6. Rough Set Results

5.3. Multilayer perceptron

For each of the three years under consideration we set out to train a backpropagation network. The topology of the networks used is: 19 neurons in the input layer, corresponding to 19 ratios, one intermediate layer whose number of neurons varied between networks, and two neurons in the output layer, corresponding to classes. The initial learning parameters also varied from network to network, as Table 7 shows.

Model	Neurons of the hidden layer	Iterations	Learning rate	Momentum
1	6	1000	0.2	0.5
2	5	1000	0.2	0.8
3	6	1000	0.5	0.7

Table 7. Multilayer Perceptron Parameters

With respect to the results obtained, Table 8 shows in percentages the correct classifications, both in the training of the networks and in their validation.

Model	Correct classifications			
	Training set		Test set	
	"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	96.3%	100%	66.67%	88.89%
	Total: 98.15%		Total: 77.78%	
2	100%	88.46%	75%	100%
	Total: 94.23%		Total: 87.5%	
3	100%	95%	100%	71.43%
	Total: 97.5%		Total: 85.71%	

Table 8. Multilayer Perceptron Results

5.4. Linear discriminant analysis

While the previous methods of classification are capable of accepting incomplete data (observations for which the value of some ratio is unknown), this is not true with Discriminant Analysis and Logistic Regression, which require that all of the data be known. If they are not known, it will be necessary to deal with the missing values in some way before performing the corresponding regressions. The first, and most conventional, alternative would be to simply eliminate those observations that have missing values. In our case, however, as little data are available, discarding an observation, which is a vector with 20 components, simply because we do not

know the value of one of those components, is not an acceptable option. In order not to lose the information provided by the known values it is a good idea to perform some type of imputation of the unknown values.

The most frequently chosen option is to substitute the unknown values with the mean or median of the known values for each variable. However, we chose for a more laborious alternative that offers more realistic imputations, described in Troyanskaya *et al*. (2001). This article compares different imputation strategies for missing values and concludes that one called KNNimpute gives the best results. For each observation with some missing value, KNNimpute looks for the nearest k observations ("k nearest neighbours") which have complete data and estimates the missing value as the mean (weighted according to the distance of the neighbour) of the corresponding values from the k nearest neighbours.

The most appropriate k value, in other words the number of neighbours of an observation that will be used to make the imputation, is determined with a procedure that consists of employing the observations without missing values. Using this information, a matrix with complete data is found. Then some values are eliminated randomly. These eliminated values are imputed taking different k values and evaluating the quality of the imputation for each k. This is done by comparing the imputed matrix with the original one. The metric used to assess the accuracy of imputation is "the Root Mean Squared difference between the imputed matrix and the original matrix, divided by the mean data value in the complete data set" (Troyanskaya *et al*., 2001). This magnitude is called normalized RMS error.

This process is repeated 50 times for each k value for the purpose of obtaining reasonable estimations. It gives a value k = 4 as the most appropriate for our data. Figure 1 records the values of the normalized RMS error for different values of k and shows how the minimum error value is reached for k = 4. Consequently, we perform an imputation of the missing values in agreement with the KNNimpute method using this value for k. With data processed in this way we can now carry out the Discriminant Analysis and the Logistic Regression.

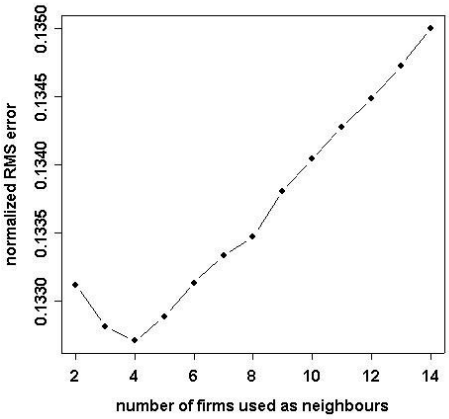

Figure 1. Error in the imputation according to the number of firms taken as neighbours to impute

Before beginning the Discriminant Analysis we did a t-test to check if the difference in mean values of the ratios between the two groups of companies was significant. The results for year 1 are shown in Appendix E, which records the mean values in each group and the p-value of the test for each ratio. It is seen that the only significant differences in means were for R1 and R9, meaning the majority of the information available does not seem, at first glance, to have a great potential to discriminate between both types of companies.

To check the verification of the normality hypothesis we carried out a univariant Shapiro-Wilk test for each ratio and each class. The results for year 1 appear in Appendix F, which contains the p-values of the test. It is observed that almost none of the ratios are distributed normally. We also assay a multivariant Shapiro-Wilk test (Table 9) and, as expected, the null hypothesis of normality was also rejected.

Shapiro-Wilk normality test data: failed firms W = 0.2, p-value = 4.429e-11	Shapiro-Wilk normality test data: healthy firms W = 0.3, p-value = 1.565e-10

Table 9. Multivariant Normality Test (Year 1)

We also checked if it was possible to accept the homoscedasticity hypothesis, that is, if the covariance matrix is the same for the two classes of companies. To do this we performed a Fligner-Killeen test that has been determined as one of the test for homogeneity of variances which is most robust against departures from normality (Conover *et al.*, 1981). This is a univariant test that contrasts the

equality of the variance between both populations for each ratio (i.e., it contrasts the equality between both populations of each element of the diagonal of the covariance matrix). The results for Year 1 are shown in Appendix G and lead us to discard the null hypothesis of equality of the variances for ratios R5, R6, R7, R8, R11, R12, R14, R20 and R21, making it necessary to reject the hypothesis of homoscedasticity.

Similar results were found for Years 2 and 3. Consequently, as the hypothesis of normality and homoscedasticity are not corroborated for any of the three years before failure, it is not possible to affirm LDA possesses an optimal classifying ability. However, the classifying ability might still be reasonably good and LDA should not be completely discarded. Thus, we can go ahead and construct the discriminant functions for each of the three years using the 19 available ratios. Next, the resulting classification with these functions is indicated in Table 10:

Model	Correct classifications			
	Training set		Test set	
	"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	85.18%	77.78%	66.67%	66.67%
	Total: 81.48%		Total: 66.67%	
2	84.62%	76.92%	50%	75%
	Total: 80.77%		Total: 62.5%	
3	85%	85%	71.43%	57.14%
	Total: 85%		Total: 64.29%	

Table 10. LDA Results Using 19 Ratios

Two additional points should be kept in mind. First, the estimation of the covariance matrix used to construct the discriminant functions is very sensitive to the presence of outliers. A simple graphic analysis of each ratio using boxplot diagrams shows a series of values (representing 2-3% of the total) that are sufficiently far from the mean to be considered outliers. Since eliminating observations that contain some atypical data would reduce the size of the sample, we opted to retain such observations and conduct the LDA through a robust estimation of the matrix of the covariances following a procedure proposed by Rousseeuw (1984) and Rousseeuw and Leroy (1987) and known as Minimum Volume Ellipsoid. Accordingly, with n observations and p variables, the procedure obtains an initial estimation of the vector of the means and the covariance matrix taken from a set of "good" observations. These good observations would be ones

considered as belonging to an ellipsoid of minimum volume that contains (n + p + 1) / 2 observations. This is refined by including those points whose Mahalanobis distance from the initial mean using the initial covariance is not too large.

Another problem is that the number of available ratios (19) is large and could hamper achieving the correct interpretation of the results. It would be wise to carry out a discrimination based on a smaller set of variables that are genuinely relevant for the classification. Also, high correlations exist between some of the variables. This leads to problems of colinearity and makes the resulting estimations unstable and very sensitive to small variations in the starting data.

For all these reasons it is advisable to conduct a previous selection of the variables that will be used in the LDA and the Logistic Regression. We chose for this selection the Akaike Information Criterion or AIC (Akaike, 1974 and 1981) which uses ideas from Information Theory to select the model that minimizes the expression: $-2\log\left[L\left(\hat{\theta}\right)\right]+2p$, where p is the number of parameters of the model (in our case the number of ratios that are included in it), $L(.)$ is the likelihood function and $\hat{\theta}$ is the maximum likelihood estimate of the parameters of the model. The AIC criterion has a significant theoretical basis and with sufficiently large sample sizes usually gives models that produce excellent classifications. However, with small-sized samples it can lead to models with too many parameters. In such cases it would be a good idea to use the Bayesian Information Criterion or BIC (Peña, 2002), which involves selecting the model that minimizes the quantity: $-2\log\left[L\left(\hat{\theta}\right)\right]+p\log n$, where n is the number of observations. This criterion penalizes most the models with higher numbers of parameters and therefore selects more parsimonious models.

In our case, the ratios selected by both criteria to perform the LDA appear in Table 11.

	AIC	BIC
Year 1	R1, R5, R8, R9, R10, R18	R5, R8, R9, R10, R18
Year 2	R2, R4, R7, R9, R10, R11, R12, R14, R20, R21	R7, R9, R10, R11, R12, R14, R20, R21
Year 3	R1, R2, R3, R4, R6, R7, R8, R9, R13, R14, R18, R19	R1, R2, R3, R4, R18[1]

Table 11. Ratios Selected by AIC and BIC Criteria to Perform LDA

Next, besides the initial model with all the ratios and the normal, non-robust estimation, the following models were generated:

- A model with all the ratios and a robust estimation.
- Two models with the ratios given by the AIC criterion, one with normal estimation and the other with robust estimation.
- Two models with the ratios given by the BIC criterion, one with normal estimation and the other with robust estimation.

In this way, for each year, six different discriminant models were available. Table 12 outlines the results produced with the best discriminant model over the test set for each one of the three years.

Model	Correct classifications			
	Training set		Test set	
	"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	85.18%	77.78%	66.67%	66.67%
	Total: 81.48%		Total: 66.67%	
2	84.62%	61.54%	62.5%	62.5%
	Total: 73.08%		Total: 62.5%	
3	85%	75%	85.71%	71.43%
	Total: 80%		Total: 78.57%	

Table 12. LDA Results

Model 1 is constructed with all the ratios and normal estimation, Model 2 with the ratios resulting from the BIC criterion and robust estimation, and Model 3 with the ratios from BIC and normal estimation. The coefficients of the discriminant

[1] In this case the penalization applied by criterion BIC to the number of parameters of the model is so high that all are eliminated. After various attempts, the ratios selected were the last five eliminated by applying the criterion, as these were the most significant ratios with respect to discrimination.

function for Year 1 appear in Appendix H. The canonical F-statistic of each model is shown in Table 13 (we reject the null hypotheses; this means that the three models are discriminant).

	Canonical F-statistic	p-value
Model 1	55.7	9.04e-10
Model 2	101.0	1.36e-13
Model 3	14.0	6.03e-4

Table 13. Canonical F-statistic of the Discriminant Models

In general, it is observed that the BIC criterion is the one that gives the best results to obtain parsimonious models. Such models will tend to present a smaller overfitting, meaning they will better classify the elements in the test set. On the other hand, robust estimation does not seem to lead to appreciable improvements in the results of the classification.

5.5. Logistic regression

The results obtained with the classification of the training and test elements by Logistic Regressions that used the 19 available ratios are shown in Table 14, for each of the three years.

Model	Correct classifications			
	Training set		Test set	
	"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	85.19%	88.89%	66.67%	66.67%
	Total: 87.04%		Total: 66.67%	
2	84.62%	92.31%	50%	75%
	Total: 88.46%		Total: 62.5%	
3	85%	85%	57.14%	42.86%
	Total: 85%		Total: 50%	

Table 14. Logistic Regression results using 19 ratios

Criteria AIC and BIC can be used to select the variables in a generalized linear model, like Logistic Regression, in the same way that was done in the case of LDA. Table 15 shows the ratios selected for each of these criteria.

	AIC	BIC
Year 1	R1, R5, R7, R8, R9, R10, R11, R13, R14, R18, R19	R5, R8, R9, R10, R18
Year 2	R1, R2, R3, R4, R7, R9, R10, R11, R14, R17, R21	R7, R9, R10, R11, R14, R21
Year 3	R2, R3, R4, R13, R14, R18, R19	R2, R3, R4, R13, R14, R18[2]

Table 15. Ratios Selected by AIC and BIC Criteria to Perform Logistic Regression

In this way, for each year we have three different logit models. Table 16 describes the results obtained with the best of the three models (judged over the test set) for each of the three years.

Model	Correct classifications			
	Training set		Test set	
	"Healthy" firms	"Failed" firms	"Healthy" firms	"Failed" firms
1	85.19%	85.19%	66.67%	66.67%
	Total: 85.19%		Total: 66.67%	
2	92.31%	88.46%	75%	87.5%
	Total: 90.38%		Total: 81.25%	
3	80%	75%	57.14%	71.43%
	Total: 77.5%		Total: 64.29%	

Table 16. Logistic Regression Results

Models 1 and 2 are obtained with the AIC criterion and Model 3 corresponds to the BIC criterion. The coefficients of *Model 1* appear in Appendix I.

5.6. Results comparison

In order to make easier the comparison between the five approaches, in Table 17 the results for the test samples are shown, in percent of correctly classified firms.

[2] Criterion BIC eliminates all the ratios. The last six eliminated were selected.

Model	Technique	Set of variables	Correct classifications	
			"Healthy" firms	"Failed" firms
1	See5 (8 leaves)	R13, R9, R17, R1, R2, R6	77.78%	66.77%
	Rough Set (27 decision rules)	R3, R4, R9, R14, R17	Total: 72.22%	
			77.78%	77.78%
	Multilayer Perceptron	All	Total: 77.78%	
			66.67%	88.89%
	LDA	All	Total: 77.78%	
			66.67%	66.67%
	Logistic Regression	R1, R5, R7, R8, R9, R10, R11,R13, R14, R18, R19	Total: 66.67%	
			66.67%	66.67%
2	See5 (6 leaves)	R1, R13, R20, R7, R3	Total: 66.67%	
			87.5%	75%
	Rough Set (25 decision rules)	R1, R3, R4, R5, R17	Total: 81.25%	
			75%	75%
	Multilayer Perceptron	All	Total: 75%	
			75%	100%
	LDA	R7, R9, R10, R11, R12, R14, R20, R21	Total: 87.5%	
			62.5%	62.5%
	Logistic Regression	R1, R2, R3, R4, R7, R9, R10, R11, R14, R17, R21	Total: 62.5%	
			75%	87.5%
3	See5 (5 leaves)	R4, R19, R1	Total: 81.25%	
			100%	57.14%
	Rough Set (22 decision rules)	R3, R4, R14, R17	Total: 78.57%	
			71.43%	57.14%
	Multilayer Perceptron	All	Total: 64.29%	
			100%	71.43%
	LDA	R1, R2, R3, R4, R18	Total: 85.71%	
			85.71%	71.43%
	Logistic Regression	R2, R3, R4, R13, R14, R18	Total: 78.57%	
			57.14%	71.43%
			Total: 64.29%	

Table 17. Results Comparison

And Table 18 displays the average accuracy of 3 years for each technique.

Technique	Correct classifications
Multilayer Perceptron	83.66%
See5	77.35%
Rough Set	72.36%
Logistic Regression	70.74%
LDA	69.25%

Table 18. Average Accuracy

Roughly speaking Multilayer Perceptron outperforms clearly the rest of the techniques, but provides non-interpretable models and, therefore, it doesn't allow knowing the relative importance of the variables to get a classification. See5 is found in second place among the better techniques in classifying and, except for year 1, outperforms the Rough Set approach. Moreover, as we could see previously, results of See5 for some models can be clearly improved by means of boosting, and could even exceed the Multilayer Perceptron. Nevertheless, we are not interested in improving the accuracy by means of losing power of explanation. If we call the boosting option, it provides models that we cannot easily understand. Then, the main advantage of See5 would be vanished, that's why we don't take that way. Furthermore, See5 provides simpler decision models than Rough Set (for example, for the year 1, See5 supplies 8 rules instead of the 27 rules provided by RS).

On the one hand, the Rough Set approach outperforms slightly the Logistic Regression and LDA; though it chooses groups of ratios far smaller than the last mentioned techniques. On the other hand, models obtained by Logistic Regression seem not to improve those provided through the LDA.

In general, machine learning techniques make a better use of the available information than statistical ones, which leads to a higher correct classification rate. Probably the structure of data space is too much complex to achieve a good classification with a linear hypersurface as LDA does it. The more sophisticated rules generated by machine learning techniques adapt better to data structure. They are very powerful tools to capture the peculiarities of data in detail.

When a model is developed, every technique uses a quite different set of variables. However, differences between models are not as great as they seem because of the correlations between the variables. If some different variables are correlated, they can provide the same information for the models.

Naturally, the ratios which appear in the solutions are not the same ones for each year, because the prediction of the insolvency achieved by each model will be one, two and three years in advance, respectively. We have considered that the ratios which appear in three of the four solutions achieved by See5, Rough Set (RS), LDA and Logistic Regression (LR), for each year, are highly discriminatory variables between solvent and insolvent firms. We refer to them as the "best ratios". Table 19 shows the ratios used inside each model and the "best ratios" in each year (except for the Multilayer Perceptron, which doesn't choose a subset of ratios).

	Year 1					Year 2					Year 3				
	See 5	RS	LDA	LR	Best Ratios	See 5	RS	LDA	LR	Best Ratios	See 5	RS	LDA	LR	Best Ratios
R1	*		*	*	*	*	*		*	*	*		*		
R2	*		*						*				*	*	
R3		*	*			*	*		*	*		*	*	*	*
R4		*	*				*		*		*	*	*	*	*
R5			*	*			*								
R6	*		*												
R7			*	*		*		*	*	*					
R8			*	*											
R9	*	*	*	*	*			*	*						
R10			*	*				*	*						
R11			*	*				*	*						
R12			*					*							
R13	*		*			*								*	
R14		*	*	*	*			*	*			*		*	
R17	*	*	*		*		*		*			*			
R18			*	*									*	*	
R19			*	*							*				
R20			*			*		*							
R21			*					*	*						

Table 19. Best Ratios

Consequently, those parts interested in evaluating the solvency of non-life insurance companies should keep in mind the following issues:

- R1- One of the most important issues in order to assure the proper functioning of any firm is to have enough liquidity. However, in the case of an insurance firm, the lack of liquidity should not arise, due to premiums are paid in before claims occur. If an insurance firm cannot pay the incurred claims, the clients and general public could lose confidence in that company. On the other hand, this ratio is a measure of financial equilibrium: if it is positive it means that the working capital is also positive.

- R3- This ratio indicates that to obtain enough financial incomes is a critical issue because nowadays these incomes are the main source of benefits for an insurance company.

- R4- This ratio is a general measure of profitability. The variable that appears in the numerator is the cashflow (cashflow plus extraordinary results) because sometimes it would be better to use this variable than profits because the first one is less manipulated than the second one. In any case, it is necessary to generate sufficient profitability to follow a right self-financing.

- R7 and R14- These are strictu sensu solvency ratios. The numerator shows the risk exposure through earned premiums (R7) or incurred claims (R14). The denominator shows the real financial support because technical provisions are considered together with capital and reserves. This demonstrates the need of having sufficient shareholder' funds and the need of complying correctly with the technical provisions to guarantee the financial viability of the insurance company.

- R9- This ratio shows what proportion of the total liabilities represent the shareholders' funds (capital and reserves). It confirms the importance, from a solvency viewpoint, of the adequacy of the mentioned funds, due to these resources could be required to meet the future claims obligations of the insurer in some eventualities.

- R17- This ratio is a traditional measuring of underwriting profitability and it indicates if the firm is following a correct rating in order to calculate right premiums that take into account the whole costs.

6. RESEARCH LIMITATIONS

This research has certain limitations that must be stated. On the one hand, the sample size is small and the number of variables is quite high. This fact produces that the predictors space is "empty" so this could increase the overfitting (this is the well-known "curse of dimensionality").

Furthermore, we have used a matched sample in order to avoid the firm size effect instead of using size as a potential predictive variable. This decision could

be questioned, but the firm size is a very important variable when we intend to forecast the business failure, especially in insurance sector, so this way we avoid putting in the shade the role of the financial variables we are interested in. Moreover, most prior research focused in Spanish insurance sector worked with matched samples (see section 2), so we can keep the comparability of our research with the previous ones.

Also it could have been desirable to carry out a jackknife validation. Yet this is nonsense for Rough Set approach due to the role that the decision maker plays in choosing the reducts. If an object of the sample is removed, the decision maker should choose a new reduct. But the new reduct chosen should be the original one unless the new one contains fewer ratios. However, this will happen in very few cases that can be considered as outliers. So we have employed a suitable validation method for all the techniques in order to compare them.

On the other hand, it would be interesting to develop an only model containing ratios from several years before bankruptcy but this fact increases the number of variables in contrast to the sample size. So we should study the ratios carefully in order to decide the more suitable ones to introduce in the model. In that sense, our research can be considered as a previous step for the construction of multi-year models.

7. CONCLUSIONS

In this paper we have compared the predictive accuracy of two data analysis methodologies of the field of Machine Learning (See5 algorithm and Rough Set methodology) on a sample of Spanish non-life insurance companies, using 19 financial ratios most of them specifically proposed for evaluating insolvency inside insurance sector. Furthermore, in order to assess the efficency of these methods, we have compared them with other widely used ones: Linear Discriminant Analysis, Logistic Regression and Multilayer Perceptron.

As shown by the experiments carried out, both machine learning approaches (See5 and Rough Set) are competitive alternatives to existing bankruptcy prediction models in insurance sector and have great potential capacities that undoubtedly make them attractive for application to the field of business classification.

Our empirical results show that these methods offer better predictive accuracy than the statistical ones that we have developed, especially the See5 algorithm. Moreover, these techniques don't require adopting restrictive assumptions about the characteristics of probability distributions of the variables and errors of the models and the decision models provided by them are easily understandable and interpretable.

In practical terms, the trees and decision rules generated could be used to preselect companies to examine more thoroughly, quickly and inexpensively, thereby, managing the financial user's time efficiently. They can also be used to check and monitor insurance firms as a "warning system" for insurance regulators, investors, management, financial analysts, banks, auditors, policy holders and consumers.

However, our work has some limitations, such as the few available cases and the uncertain quality of some information. Furthermore, if we want to use these models for predicting insolvency, we should keep in mind that they have been developed without including some aspects which could be relevant for this issue, such as size and industry.

But in spite of these problems, our focus is to show the suitability of these machine learning techniques as support decision methods for insurance sector. In short, we believe that these methods, without replacing analyst's opinion and in combination with other ones, will play a bright role in the decision-making process inside insurance sector.

8. REFERENCES

AKAIKE, H. (1974): "A new look at statistical model identification", *IEEE Transactions on Automatic Control*, 19, 716-723.

AKAIKE, H. (1981): "Likelihood of a model and information criteria", *Journal of Econometrics*, 16, 3-14.

ALTMAN, E.I. (1968): "Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy", *Journal of Finance*, 23 (4), 589-609.

ALTMAN, E.I.; HALDEMAN, R.G.; NARAYANAN, P. (1977): "ZETA™ analysis: a new model to identify bankruptcy risk of corporations", *Journal of Banking and Finance*, 1 (1), 29-54.

ALTMAN, E.I.; MARCO, G.; VARETTO, F. (1994): "Corporate distress diagnosis: comparisions using linear discriminant analysis and neural networks (the Italian experience)", *Journal of Banking and Finance*, 18 (3), 505-529.

AMBROSE, J.M.; CARROLL, A.M. (1994): "Using Best's Ratings in Life Insurer Insolvency Prediction", *The Journal of Risk and Insurance*, 61 (2), 317-327.

BAR-NIV, R.; SMITH, M.L. (1987): "Underwriting, Investment and Solvency", *Journal of Insurance Regulation*, 5, 409-428.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. (1984): *Classification and regression trees*, Wadsworth, Belmont.

CLARK, P.; BOSWELL, R. (1991): "Rule Induction with CN2: Some Recent Improvements", in KODRATOFF, Y. (Ed.): *Machine Learning - Proceedings of the Fifth European Conference (EWSL-91)*, Springer-Verlag, Berlin, 151-163.

CONOVER, W.J.; JOHNSON, M.E.; JOHNSON, M.M. (1981): "A comparative study of test for homogeneity of variances, with applications to the outer continental shelf bidding data", *Technometrics*, 23, 351-361.

DE ANDRÉS, J. (2001): "Statistical Techniques vs. SEE5 Algorithm. An Application to a Small Business Environment", *The International Journal of Digital Accounting Research*, 1 (2), 153-179.

DIMITRAS, A.I.; SLOWINSKI, R.; SUSMAGA, R.; ZOPOUNIDIS, C. (1999): "Business failure prediction using Rough Sets", *European Journal of Operational Research,* 114, 263-280.

DIZDAREVIC, S.; LARRAÑAGA, P.; PEÑA, J.M.; SIERRA, B.; GALLEGO, M.J.; LOZANO, J.A. (1999): "Predicción del fracaso empresarial mediante la combinación de clasificadores provenientes de la estadística y el aprendizaje automático", in Bonsón, E. (Ed.): *Tecnologías Inteligentes para la Gestión Empresarial*, RA-MA Editorial, Madrid, 71-113.

FREUND, Y.; SCHAPIRE, R.E. (1997): "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55(1), 119-139.

GRECO, S.; MATARAZZO, B.; SLOWINSKI, R. (1998): "A new rough set approach to evaluation of bankruptcy risk", in ZOPOUNIDIS, C. (Ed.): *New Operational Tools in the Management of Financial Risks*, Kluwer Academic Publishers, Dordrecht, 121-136.

HENRICHON, Jr.; E.G.; FU, K.S. (1969): "A nonparametric partitioning procedure for pattern classification", *IEEE Transactions on Computers*, 18, 614-624.

LÓPEZ, D.; MORENO, J.; RODRÍGUEZ, P. (1994): "Modelos de previsión del fracaso empresarial: Aplicación a entidades de seguros en España", *Esic Market*, 84, 83-125.

MARTÍN, M.L.; LEGUEY, S.; SÁNCHEZ, J. M. (1999): *Solvencia y estabilidad financiera en la empresa de seguros: Metodología y evaluación empírica mediante análisis multivariante*. Cuadernos de la Fundación Mapfre Estudios, nº 49, Madrid.

MARTÍNEZ DE LEJARZA, I. (1999): "Previsión del fracaso empresarial mediante redes neuronales: un estudio comparativo con el análisis discriminante", in Bonsón, E. (Ed.): *Tecnologías Inteligentes para la Gestión Empresarial*, RA-MA Editorial, Madrid, 53-70.

MCKEE, T.E.; LENSBERG, T. (2002): "Genetic programming and rough sets: a hybrid approach to bankruptcy classification", *European Journal of Operational Research*, 138, 436-451.

MORA, A. (1994): "Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit", *Revista Española de Financiación y Contabilidad*, 23 (78), 203-233.

MORGAN, J.N.; MESSENGER, R.C. (1973): *THAID: a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables*, Survey Research Center, Institute for Social Research, University of Michigan.

MORGAN, J.N.; SONQUIST, J.A. (1963): "Problems in the analysis of survey data, and a proposal", *Journal of the American Statistical Association*, 58, 415-434.

MÜLLER GROUP (1997): *Müller Group Report. 1997. Solvency of insurance undertakings*, Conference of Insurance Supervisory Authorities of The Member States of The European Union.

NIBLETT, T. (1987): "Constructing decision trees in noisy domains", in BRATKO, I.; LAVRA , N. (Eds.): *Progress in Machine Learning (proceedings of the 2nd European Working Session on Learning)*, Sigma, Wilmslow, UK, 67-78.

NURMI, H.; KACPRZYK, J.; FEDRIZZI, M. (1996): "Probabilistic, fuzzy and rough concepts in social choice", *European Journal of Operational Research*, 95, 264-277.

PAWLAK, Z. (1991): *Rough Sets. Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht/ Boston/ London.

PEÑA, D. (2002): *Análisis de datos multivariantes*, McGraw-Hill, Madrid.

PREDKI, B.; SLOWINSKI, R.; STEFANOWSKI, J.; SUSMAGA, R.; WILK, S. (1998): "ROSE – Software Implementation of the Rough Set Theory", in POLKOWSKI, L.; SKOWRON, A. (Eds.): *Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence,* 1424, Springer-Verlag, Berlin, 605-608.

PREDKI, B.; WILK, S. (1999): "Rough Set Based Data Exploration Using ROSE System", in RAS, Z.W.; SKOWRON, A. (Eds.): *Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence,* 1609, Springer-Verlag, Berlin, 172-180.

QUINLAN, J.R. (1979): "Discovering rules by induction from large collections of examples", in Michie, D. (Ed.): *Expert systems in the microelectronic age*, Edimburgh University Press, Edimburgh.

QUINLAN, J.R. (1983): "Learning efficient classification procedures", in *Machine learning: an Artificial Intelligence approach*, Tioga Press, Palo Alto.

QUINLAN, J.R. (1986): "Induction of decision trees", *Machine Learning*, 1 (1), 81-106.

QUINLAN, J.R. (1988): "Decision trees and multivalued attributes", *Machine Intelligence*, 11, 305-318.

QUINLAN, J.R. (1993): *C4.5: Programs for machine learning*, Morgan Kaufmann, California.

QUINLAN, J.R. (1997): *See5* (URL http://www.rulequest.com/see5-info.html).

R DEVELOPMENT CORE TEAM (2004): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (URL http://www.R-project.org).

REZA, F.M. (1994): *An introduction to Information Theory*, Dover Publications, New York.

ROSENBLATT, F. (1957): "The perceptron: A perceiving and recognizing automation", Technical Report 85-460-1, Cornell Aeronautical Laboratory.

ROUSSEEUW, P.J. (1984): "Least median of squares regression", *Journal of the American Statistical Association*, 79, 871-881.

ROUSSEEUW, P.J.; LEROY, A.M. (1987): *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.

RUMELHART, D.; HINTON, G.; WILLIAMS, R. (1986): "Learning representations by back-propagating errors", *Nature*, 323, 533-536.

SALCEDO, S.; FERNÁNDEZ, J.L.; SEGOVIA, M.J.; BOUSOÑO, C. (2005): "Genetic Programming for the Prediction of Insolvency in Non-life Insurance Companies", *Computers & Operations Research*, 32 (4), 749-765.

SANCHÍS, A.; GIL, J.A.; HERAS, A. (2003): "El análisis discriminante en la previsión de la insolvencia en las empresas de seguros no vida", *Revista Española de Financiación y Contabilidad*, 32 (116), 183-233.

SEGOVIA, M.J.; SALCEDO, S.; BOUSOÑO, C. (2004): "Prediction of Insolvency in Non-life Insurance Companies using Support Vector Machines, Genetic Algorithms and Simulated Annealing", *Fuzzy Economic Review*, 9 (1), 79-94.

SERRANO, C.; MARTÍN, B. (1993): "Predicción de la crisis bancaria mediante el empleo de redes neuronales artificiales", *Revista Española de Financiación y Contabilidad*, 22 (74), 153-176.

SETHI, I.K.; SARVARAYUDU, G.P.R. (1982): "Hierarchical classifier design using mutual information", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4, 441-445.

SLOWINSKI, R. (1993): "Rough set learning of preferential attitude in multicriteria decision making", in KOMOROWSKI, J.; RAS, Z.W. (Eds.), *Methodologies for Intelligent Systems. Lecture Notes in Artificial Intelligence*, 689, Springer-Verlag, Berlin, 642-651.

TAM, K.Y. (1991): "Neural network models and the prediction of bankruptcy", *Omega*, 19 (5), 429-445.

TROYANSKAYA, O.; CANTOR, M.; SHERLOCK, G.; BROWN, P.; HASTIE, T.; TIBSHIRANI, R.; BOTSTEIN, D.; ALTMAN, R.B. (2001): "Missing value estimation methods for DNA microarrays", *Bioinformatics*, 17(6), 520-525.

WEBB, A. (2002): *Statistical Pattern Recognition*, John Wiley and Sons, Chichester.

WIDROW, B.; HOFF, M.E. (1960): "Adaptive switching circuits", Institute of Radio Engineers, *Western Electronic Show and Convention*, Convention Record, part 4, 96-104.

WITTEN, I.H.; FRANK, E. (2000): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, California.

Appendix A: List of Ratios

RATIO	DEFINITION
R1	Working capital/ Total Assets
R2	Earnings before Taxes (EBT)/ (Capital+ Reserves)
R3	Investment Income/ Investments
R4	EBT*/ Total Liabilities EBT* = EBT+ Reserves for Depreciation+ Provisions + (Extraordinary Income-Extraordinary Charges)
R5	Earned Premiums/ (Capital+ Reserves)
R6	Earned Premiums Net of Reinsurance/ (Capital+ Reserves)
R7	Earned Premiums/ (Capital+ Reserves+ Technical Provisions)
R8	Earned Premiums Net of Reinsurance/ (Capital+ Reserves+ Technical Provisions)
R9	(Capital +Reserves)/ Total Liabilities
R10	Technical Provisions/ (Capital + Reserves)
R11	Claims Incurred/ (Capital+ Reserves)
R12	Claims Incurred Net of Reinsurance/ (Capital+ Reserves)
R13	Claims Incurred / (Capital+ Reserves + Technical Provisions)
R14	Claims Incurred Net of Reinsurance/ (Capital+ Reserves+ Technical provisions)
R15	(Claims Incurred/ Earned Premiums)+ (Other Charges and Commissions/ Other Income)
R16	(Claims Incurred Net of Reinsurance/ Earned Premiums Net of Reinsurance)+ (Other Charges and Commissions/ Other Income)
R17	(Claims Incurred + Other Charges and Commissions)/ Earned Premiums
R18	(Claims Incurred Net of Reinsurance + Other Charges and Commissions)/ Earned Premiums Net of Reinsurance
R19	Technical Provisions of Assigned Reinsurance/ Technical Provisions
R20	Claims Incurred / Earned Premiums
R21	Claims Incurred Net of Reinsurance / Earned Premiums Net of Reinsurance

Appendix B: List of Subintervals (quartiles) (Rough Set approach)

Ratio	1st	2nd	3rd	4th
R1	(-∞, 0.115]	(0.115, 0.295]	(0.295, 0.475]	(0.475, +∞)
R2	(-∞, 0]	(0, 0.1]	(0.1, 0.07]	(0.07, +∞)
R3	(-∞, 0.03]	(0.03, 0.06]	(0.06, 0.11]	(0.11, +∞)
R4	(-∞, 0.03]	(0.03, 0.08]	(0.08, 0.26]	(0.26, +∞)
R5	(-∞, 0.565]	(0.565, 1.565]	(1.565, 3.29]	(3.29, +∞)
R6	(-∞, 0.525]	(0.525, 1.38]	(1.38, 2.715]	(2.715, +∞)
R7	(-∞, 0.455]	(0.455, 0.725]	(0.725, 1.22]	(1.22, +∞)
R8	(-∞, 0.46]	(0.46, 0.7]	(0.7, 1.18]	(1.18, +∞)
R9	(-∞, 0.14]	(0.14, 0.35]	(0.35, 0.68]	(0.68, +∞)
R10	(-∞, 0.04]	(0.04, 0.545]	(0.545, 2.97]	(2.97, +∞)
R11	(-∞, 0.27]	(0.27, 1.095]	(1.095, 2.43]	(2.43, +∞)
R12	(-∞, 0.27]	(0.27, 0.845]	(0.845, 1.815]	(1.815, +∞)
R13	(-∞, 0.27]	(0.27, 0.49]	(0.49, 0.82]	(0.82, +∞)
R14	(-∞, 0.225]	(0.225, 0.435]	(0.435, 0.765]	(0.765, +∞)
R17	(-∞, 0.98]	(0.98, 1.055]	(1.055, 1.27]	(1.27, +∞)
R18	(-∞, 1]	(1, 1.09]	(1.09, 1.29]	(1.29, +∞)
R19	(-∞, 0]	(0, 0.065]	(0.065, 0.19]	(0.19, +∞)
R20	(-∞, 0.515]	(0.515, 0.68]	(0.68, 0.785]	(0.785, +∞)
R21	(-∞,0.515]	(0.515, 0.655]	(0.655, 0.75]	(0.75, +∞)

Appendix C: Assignment of codes to subintervals (Rough Set approach)

Ratio	1st	2nd	3rd	4th
R1	1	2	3	4
R2	1	2	3	4
R3	1	2	3	4
R4	1	2	3	4
R5	1	3	4	2
R6	1	3	4	2
R7	1	3	4	2
R8	1	3	4	2
R9	1	3	4	2
R10	1	3	4	2
R11	1	4	3	2
R12	1	4	3	2
R13	1	4	3	2
R14	1	4	3	2
R17	1	4	3	2
R18	1	4	3	2
R19	1	3	3	2
R20	4	3	2	1
R21	4	3	2	1

Appendix D: The 27 rules algorithm – Model 1 (Rough Set approach)

Rule number	R3	R4	R9	R14	R17	Decision 0 = failed 1= healthy	Strength	Firms
1	2		4			0	4	F2, F14, F18, F32
2		3		2		0	4	F17, F30, F43, F35
3	1	4				0	3	F7, F10, F31
4		1	4			0	3	F13, F28, F33
5			3	3		0	2	F1, F4
6	3		1			0	2	F6, F8
7	2		3	4		0	2	F11, F15
8	4			1		0	2	F12, F16
9	4	1				0	2	F16, F29
10	2		1	3		0	1	F9
11			3	2		0	1	F3
12	3		2		2	0	1	F5
13	2	2		2		0	3	F2, F18, F36
14				1	3	1	3	F102, F113, F132
15			2		4	1	5	F114, F117, F131, F133, F134
16		1	3			1	4	F101, F109, F111, F115
17	4	2				1	2	F106, F110
18	3	4			1	1	2	F103, F105
19	1	2	4			1	2	F112, F129
20		2			2	1	1	F135
21		3	1			1	1	F104
22	2			1		1	2	F116, F130
23	3			4		1	3	F107, F111, F115
24	4			3		1	2	F106, F108
25	1	3	2			1	1	F128
26	1	1			2	1	1	F136
27		4		2	4	1	1	F118

Appendix E: t-test of equality of means (year 1)

Ratio	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
p-value	0.00	0.97	0.96	0.19	0.56	0.7	0.16	0.26	0.05	0.09
Mean healthy	0.41	0.08	0.10	0.12	1.98	1.6	0.71	0.69	0.48	2.09
Mean failed	0.13	0.08	0.10	0.20	1.39	1.3	3.54	2.98	0.25	0.56
Ratio	**R11**	**R12**	**R13**	**R14**	**R17**	**R18**	**R19**	**R20**	**R21**	
p-value	0.38	0.46	0.15	0.27	0.36	0.35	0.53	1.00	0.91	
Mean healthy	1.51	1.19	0.48	0.46	1.15	1.21	0.16	0.65	0.64	
Mean failed	0.88	0.71	2.34	1.86	1.45	1.51	0.12	0.65	0.63	

Appendix F: p-values for the univariate Shapiro-Wilk test for normality (year 1)

Ratio	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
healthy	0.9	9e-07	4e-08	8e-04	0.002	0.004	7e-05	6e-05	0.003	3e-04
failed	0.5	1e-09	4e-07	0.001	0.087	0.131	3e-10	4e-10	0.016	7e-03

Ratio	**R11**	**R12**	**R13**	**R14**	**R17**	**R18**	**R19**	**R20**	**R21**	
healthy	6e-04	7e-04	7e-05	3e-05	9e-03	8e-03	5e-05	0.6	0.9	
failed	5e-02	3e-02	5e-10	5e-10	2e-10	5e-10	1e-06	4e-04	2e-04	

Appendix G: p-values of the test for homogeneity of variances (year 1)

Ratio	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
p-value	0.19	0.11	0.28	0.12	0.002	0.00067	0.049	0.048	0.36	0.77

Ratio	R11	R12	R13	R14	R17	R18	R19	R20	R21
p-value	0.014	0.0035	0.074	0.044	0.7	0.88	0.13	0.012	0.029

Appendix H: Coefficients of the discriminant functions (year 1)

Ratio	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Coefficient	-1.12	-1.22	0.22	0.85	3.52	-2.5	-33.9	32.60	-2.2	-0.66

Ratio	R11	R12	R13	R14	R17	R18	R19	R20	R21	
Coefficient	-4.66	4.13	41	-38.39	-2.9	3.14	1.90	-8.3	6.6	

Appendix I: Coefficients of the logistic regression (year 1)

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.5592	2.5680	-0.997	0.3190
R1	-2.5135	1.4931	-1.683	0.0923
R5	4.9185	2.3843	2.063	0.0391
R7	-19.9400	13.1296	-1.519	0.1288
R8	16.1611	10.6370	1.519	0.1287
R9	-3.7963	2.6070	-1.456	0.1453
R10	-1.4987	0.6919	-2.166	0.0303
R11	-5.0404	3.5154	-1.434	0.1516
R13	24.3985	16.0958	1.516	0.1296
R14	-17.7901	12.1082	-1.469	0.1418
R18	2.7515	1.7777	1.548	0.1217
R19	3.2931	2.3088	1.426	0.1538

Null deviance: 74.860 on 53 degrees of freedom

Residual deviance: 37.541 on 42 degrees of freedom