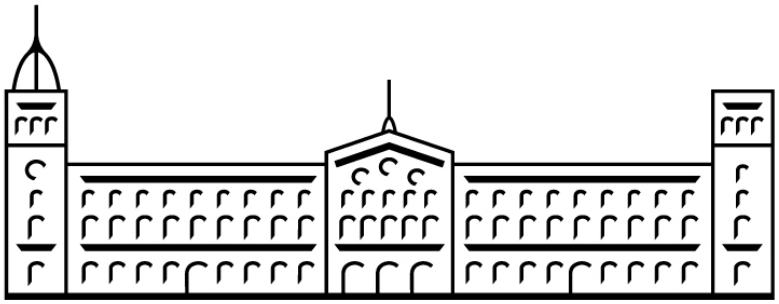

Bachelor Thesis

DOUBLE DEGREE IN

MATHEMATICS - BUSINESS ADMINISTRATION AND MANAGEMENT

Facultat de Matemàtiques - Facultat d'Economia i Empresa
Universitat de Barcelona

INSOLVENCY RISK: CHARACTERISATION AND PREDICTION

Author: Adrià Xaus Pariente

Tutors: Dr. Josep Fortiana Gregori - Dr. Jordi Martí Pidelaserra

Departament de Probabilitat, Lògica i Estadística - Departament de Comptabilitat

Barcelona, January 2016

Abstract

The present document sets out to analyse the concept of insolvency risk in a firm and how it can be objectively measured. Our main objective is to predict whether a firm will face an insolvency situation, based on its most recent historical data stored in its accounts.
In order to achieve it, the prediction of insolvency risk is studied reviewing some of the most relevant literature and explaining the accounting and financial implications which lie behind it, understanding the concept of insolvency from this perspective. In mathematical terms, this is an example of the so-called Problem of Classification (or Discriminant Analysis), which is usually approached using Statistics. More specifically, the chosen way to mathematically measure insolvency risk is through some of the most popular statistical prediction methods which deal with this problem. Some of these methods consist of the classical Altman's Z Score, essentially equivalent to the Linear Discriminant, or more contemporary methods like Classification and Regression Trees or Neural Networks.
These methods are applied on two samples. The first one is a sample of 40 Spanish firms selected under some certain criteria, gathering its data from SABI database (*Sistema de Análisis de Balances Ibéricos*). The second one is the sample that Professor E. I. Altman used in his famous 1968 article, where he introduced its aforementioned Z Score.
A balanced approach between financial theory and statistical theory is used in order to effectively convey the message that we cannot totally rely on the statistical methods without taking into account the non-mathematical implications, for this is a complex issue involving many other areas such as finance, accounting or economics.

Resum

El present treball té per objectiu analitzar el concepte de risc d'insolvència en una empresa i com pot ser mesurat de forma objectiva. L'objectiu principal que ens plantegem és tractar de predir si una empresa es veurà abocada a una situació d'insolvència, basant-nos en les seves dades històriques més recents a nivell comptable.
Per tal d'assolir aquest objectiu, s'estudia la predicció del risc d'insolvència repassant l'evolució històrica de la seva recerca a través de la literatura més destacada, i explicant les implicacions comptables i financeres que hi ha al seu darrere, entenent el concepte d'insolvència des d'aquest punt de vista.
En termes matemàtics, es tracta d'estudiar un exemple dins de l'anomenat Problema de Classificació (o Anàlisi Discriminant), que s'acostuma a tractar en Estadística. Concretament, la manera escollida de mesurar matemàticament el risc d'insolvència és a través d'alguns dels mètodes estadístics de predicció més populars enfocats a aquest problema. Alguns d'aquests mètodes comprenen des de la clàssica Z d'Altman, essencialment equivalent al Discriminador Lineal, fins a mètodes més actuals com són els Arbres de Classificació i Regressió o les Xarxes Neuronals.
Aquests mètodes s'apliquen a dues mostres de dades. La primera és una mostra de 40 empreses espanyoles seleccionades segons uns certs criteris, les dades de la qual han estat extretes de la base de dades SABI (*Sistema de Análisis de Balances Ibéricos*). La segona és la mostra que el Professor E. I. Altman va utilitzar al seu famós article de 1968, on va introduir l'esmentada Z d'Altman.
Hem optat per atorgar una importància similar tant a la part estrictament matemàtica com a la part estrictament no matemàtica per tal de transmetre la idea de que cal tenir molt en compte tota la teoria i totes les implicacions que queden al marge dels mètodes estadístics pròpiament dits, ja que es tracta d'un tema força complex que abraça moltes àrees tals com les finances, la comptabilitat o l'economia.

Acknowledgments

First and foremost, I wish to acknowledge the regular and caring support received from my tutors during my writing of this Bachelor Thesis. Their guidance and help was crucial to successfully understand what I was studying and complete this document. Therefore, I would like to share with them the result and make it clear that its flaws are my own responsibility.

I wish to thank Professor E. I. Altman for personally providing my tutors with the original dataset of his 1968 article, so that I could reproduce a slice of his great work.

I also wish to thank Professor Román Abades for giving one of my tutors, Dr. Josep Fortiana, some insights into SABI database.

Last but not least, I highly appreciate the endless patience of my parents. The elaboration of this document was also possible thanks to them.

Contents

1 Introduction

1.1 Motivations and Objectives

A crucial issue within the business world is the study of **corporate performance**, which is the analysis of the evolution of a firm throughout a certain period of time. Its study is of vital importance because it diagnoses the situation of a firm in a specific moment in time and gives answers to whether a firm is accomplishing its objectives. Taking into consideration the recent global financial crisis as an example, it is in periods like that when corporate performance is seriously threatened and therefore a great deal of analysis and prediction is needed.

The study of corporate performance, in its broader sense, has been largely studied since the second half of the twentieth century and in many ways. For instance, it can be based on the firm's historical data, using the so-called **Financial Statement Analysis**. However, there is still no universally-established approach capable of solving all the issues involved. Our belief is that this limitation is due to its **complexity** and **wideness**. Corporate performance is complex because there is no straight-forward way to objectively analyse it or measure it. This is because a firm's environment is usually large and therefore a lot of sources continuously contribute to the determination of its performance. Corporate performance is also wide because it involves different areas, basically Accounting, Finance, Economics and even Sociology. All of them are capable of providing arguments to its interpretation.
Moreover, as the amount of mathematical techniques (mainly statistical techniques) that are of use for such purposes and their sophistication increases, there is the need to continuously test them in order to assess their validity. Therefore, corporate performance can also be interpreted in terms of Statistics.

Consequently, for want of a better methodology, the usual way to approach its study is with a set of complementary analyses and techniques to try to cover and explain as much of it as possible, resulting in a varied list of conclusions extracted from all of them. The main objective *should* be to make sure that the results are as coherent as possible.

To deal with corporate performance as a whole is far beyond our purposes. It is too ambitious. Instead, we focus on one aspect: the **characterisation and prediction of insolvency risk**. It is of paramount importance to consider the insolvency risk of a firm because it clearly has enormous implications not only for its own becoming but also for the general interest, so it is a key point within corporate performance.

The prediction of insolvency risk using Statistics became a popular area of research with the publication of an article by E. I. Altman in 1968, where he introduced the famous Z Score. Due to the relevance of this article, it is a good starting point within our own study and it is indeed a fundamental reference.

There are two main objectives in our study. On the one hand, to introduce, from a theoretical perspective, the traits of insolvency risk: what does it mean, what does it involve and what elements can be used to its determination. On the other hand, to formally explain some statistical methods used for objectively measuring such risk. More specifically, within Statistical Theory we can apply the measure of insolvency risk as an example of a more general problem called the **Problem of Classification**. The appropriate statistical methods to this problem attempt to classify a given observation into one group (and only one) of a set of mutually-exclusive groups. In our specific case of insolvency risk there are two groups of firms in which a firm can be classified: 'failed' firms and 'non-failed' firms. The idea is that, if the prediction method classifies a firm in the 'failed' group, we expect its insolvency risk to be high, and if the prediction method classifies it in the 'non-failed' group, we expect its insolvency risk to be low.

As previously mentioned, there are numerous methods and therefore we just touch on the most popular ones: Linear Discriminant Analysis, Logistic Regression, k-Nearest Neighbours, Classification and Regression Trees, and Neural Networks.

Another objective is to test these methods using two different samples, a recent one (called Sample 1) and a historical one (called Sample 2 or Altman's Sample), so that we can assess to what extend each method predicts insolvency risk (the prediction capability) and also compare them. More specifically, we test all the methods with Sample 1 and only test Linear Discriminant Analysis with Sample 2, because this is the sample which Altman used in order to test LDA in his article.

The reason why the two complementary groups are called 'failed' and 'non-failed' is because the samples consist of firms which we already know whether they became insolvent or not. Strictly speaking though, it would be more correct, given our purposes, to separate between 'solvent' and 'insolvent' firms, or even more correct, to separate between 'firms with high insolvency risk' and 'firms with low insolvency risk'. We deal with all these terms in following sections.

In conclusion, introducing insolvency risk in this way will give us enough ground to make relevant contributions to the subject and, at the same time, explore some relations and implications involved not only in insolvency risk and in corporate performance, but also in Accounting, Finance and Statistics.
This is why this document does not aim to offer a complete self-contained view of the subject but a well-structured and coherent introduction to it, giving both the author and the reader the opportunity to develop further research from it in any of the different perspectives contained.

1.2 Methodological Approach and Structure

In order to study the characterisation and prediction of insolvency risk, in line with 1.1, we have committed ourselves to **strike a balance between Statistical Theory and Finance-Accounting Theory**, and also to **strike a balance between theory and practice**.
This is because we strongly believe that, in order to effectively present a good introduction to the subject, we cannot attach more importance to a certain part and leave aside the rest of them. In other words, *in theory*, every part should have the same relevance to the whole issue and only by studying everything proportionately we will be able to reach coherent conclusions. *In practice*, we must admit that Statistics account for a greater proportion but this does not mean at all that the approach to the subject in terms of non-statistical analysis is poor. Neither does it mean that we cannot reach coherent conclusions. They are simply different approaches and we actually consider both.
Therefore, one approach is only based on Finance and Accounting Theory, and the other approach is only based on Statistical Theory. However, we will eventually realise that there are important connections.

All things considered, the document is divided in two main sections: a theoretical framework and a practical framework.
The theoretical framework contains all the theory needed prior to testing the statistical methods, namely Finance-Accounting Theory and Statistical Theory. Firstly, we take a general look at some of the existent literature to explain how insolvency risk has been historically approached, highlighting two prominent researchers: R. A. Fisher and E. I. Altman. Secondly, we introduce all the important elements with regards to Financial and Accounting Theory. This is where key concepts such as return, solvency and risk appear. Finally, insolvency risk is treated from a mathematical perspective, explaining the mathematical idea behind it and introducing a selection of statistical methods: Linear Discriminant Analysis, Logistic Regression, k-Nearest Neighbours, Classification and Regression Trees, and Neural Networks.
The practical framework introduces the two chosen samples and the selected predictors for each of them, and explains the procedure followed to select them. Once these elements have been properly introduced, we reproduce the results of the tests, relating them to the previous section and explaining the most relevant details. The full code used in order to test the methods is included in Appendix A so that the interested reader may consult it and complement it with this section.
The document finishes with the conclusions obtained from both frameworks, specially after testing the methods, and a list of some possible further developments from which continue our work.

2 Theoretical Framework

2.1 Historical Background

2.1.1 R. A. Fisher and E. I. Altman's Pioneering Research

As stated in the introduction, corporate performance has been largely studied since the second half of the twentieth century. Nevertheless, research concerning Statistics and, more specifically the Problem of Classification, started earlier.

During the first half of the twentieth century, **Sir Ronald A. Fisher** (1890-1962) laid the foundations of **Modern Statistics** and **Experimental Design**. According to [5], not only was he pioneer in Multivariate Analysis but in almost every aspect of Statistics. He also made great contributions to Biology (giving rise to Biostatistics) and Genetics.
Regarding the Problem of Classification, the genuine article was *The Use of Multiple Measurements in Taxonomic Problems*, published in the *Annals of Eugenics* in 1936 ([9]). There, having two or more populations measured in several characters, his objective was to determine certain linear functions of the measurements by which the populations were best **discriminated**. He introduced discriminant functions using observations from three kinds (groups) of Iris flowers which became famous in time. So we can see that, for want of a more accurate explanation, the only difference between Fisher's research and our purposes is the study subject: flowers instead of firms.
The ideas that Fisher introduced in the article have evolved to what it is known as **Linear Discriminant Analysis (LDA)** or **Fisher-LDA**. We deal with it in 2.3.3, based on [8], [16] and [7]. It is also interesting the appearance in the article of the so-called **Analysis of Variance (ANOVA)**, another important statistical concept which we are not going to discuss.

In terms of corporate performance, two of the pioneering researchers to take part in its study at the beginning of the second half of the twentieth century were **William H. Beaver** (1940,-) and **Edward I. Altman** (1941,-). The former was pioneer in determining which were the appropriate variables to include in a suitable model, using Financial Statement Analysis and accounting ratios. The latter was pioneer in choosing an appropriate statistical method, publishing it in the article *Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy*, in the *Journal of Finance* in 1968 ([3]).

Altman's article has been quoted numberless times and it is one of our principal references. Its title is self-explanatory; it focuses on predicting whether a firm will go bankrupt using financial ratios and Discriminant Analysis, just like Fisher but talking about firms.
To him, the question became which ratios were most important in detecting potential bankruptcy, what weights should be attached to those selected ratios and how should the weights be objectively established. He went on to say that he chose Discriminant Analysis as the appropriate statistical technique. With regards to the ratios, he built a list of twenty-two potentially helpful ratios, again by using Financial Statement Analysis, and then he compiled it for evaluation, ending up with only five ratios which he would use as predictors. This contrasts with Beaver, who ended up selecting only one. This is why some references distinguish between **Simple Discriminant Analysis** and **Multiple Discriminant Analysis**.

Once Altman had selected the predictors, he applied the method using a sample of 66 American corporations. Our so-called Sample 2 is precisely this sample, provided by Professor Altman himself, so we test test Linear Discriminant Analysis using it. In this case, the question is whether we obtain the same results that Altman did forty-eight years ago.

If we compare Altman's article to our purposes we have to bear in mind that LDA is only one of the methods that we study and that, much as we use data regarding 'failed' and 'non-failed' firms, insolvency risk and corporate bankruptcy are *not* synonymous. We clearly explain these questions in following sections.
With respect to financial ratios, they also are the set of predictors used for Sample 1, so we must explain its specific selection. However, the ratios that we consider differ from those of Altman.

2.1.2 Posterior Research

After the research led by Beaver (1966) and Altman (1968), accounting and finance academics continued to analyse corporate performance and the prediction of corporate bankruptcy. According to [13], these analyses take two different approaches. The first one is based on **accounting data**, just like Beaver and Altman. The second one has an added source of information which is **market data**, namely information which is external to the firm and therefore it is determined by a firm's environment. Broadly speaking, we may refer to it as *the market*. Theoretically, adding market data to historical data like that stored on a firm's accounts should result in better approximations. The thing is, market data might not be always available, it rather depends on the firm's nature. On the other hand, a firm must always have some sort of accounting information available.
In the following lines we will restrict ourselves to the first approach because ours is only based on accounting data.

So, as we already know, the accounting approach basically uses financial ratios as predictors (the variables of the model). As statistical techniques became more advanced, researchers were allowed to came up with new models with greater prediction capability.

In 1977, nine years after Altman's article, he alongside **Robert G. Haldeman** and **Paul Narayanan** decided to make a second version to include seven more ratios to his model. Actually, what is relevant here is the technique used and not so much how many ratios were used (or which ratios specifically).
Another statistical method included in this document is Logistic Regression. It was first developed by **Ohlson** in 1980 and **Zmijewski** in 1984.
A Recurrent Partitioning Algorithm (RPA) was used by **Frydman**, Altman and **Kao** in 1985, whereas **Mar-Molinero and Ezzamel** in 1991, and **Mar-Molinero and Neophytou** in 2004 used a technique called Multi-Dimensional Scaling (MDS).
Neural Networks are a set of much more modern models that we decide to include in this document because of its potential. These models were first used to this purpose by **Tam and Kiang** in 1992.
Pawlak in 1982, **McKee** in 1995 and **Slowinski and Zopounidis** also in 1995 used Rough Sets.
Zopounidis and Doumpos in 1999 and **Voulgarirs, Doumpos and Zopounidis** in 2000 used Utilities Additives Discriminant (UTADIS).

Logically, the above information must not be regarded as complete. We have only presented a brief list of researchers based on [13]. In [11], for example, there is more information and the interested reader may well want to take a look at it to complete this part.
Nevertheless, we highlight a couple of interesting comments from this reference. Firstly, that models of prediction have become more and more sophisticated in order to be able to incorporate the effects of financial crises or other outstanding business episodes. Secondly, that although Neural Networks outperform traditional methods such as linear methods or heuristic methods based on simple rules of thumb, the general comparison still remains as an open question.

It is not our purpose to explain all of the above methods since not all of them are included in the document. The ones which will be duly studied in following sections are Linear Discriminant Analysis, Logistic Regression and Neural Networks.

2.2 Financial and Accounting Theory

2.2.1 Introduction to Accounting and Financial Statement Analysis

Accounting could be roughly described as the social science that aims to provide useful information about a firm to all of its **stakeholders**, namely any kind of economic agent interested in it. Usually, stakeholders are divided into the following six groups: **credit grantors**, **equity investors**, **managers**, **auditors**, **analysts** and other interested groups. We refer to [6] for its definitions.

The kind of information a stakeholder would expect about a firm is that which indicates, as thorough and clear and relevant as possible, the firm's performance over a specific period of time (a whole year, for example), so that he or she can evaluate to what extent his or her investment in the firm has been successful. Therefore, Accounting can be used (and it *is* actually used) in order to measure corporate performance.

Despite the lack of general consensus among the literature about a universal definition, what it is clear is that the main objective of Accounting can be achieved through the elaboration of the so-called **financial statements** (or **annual accounts**), a set of written documents that are subjected to a certain regulation, and therefore firms are committed to them.

In Spain, the main legal accounting document is the **Accounting General Plan** (***Plan General de Contabilidad*** ([15])). In its first part, it states that a firm's financial statements consist of five documents, being the **balance sheet** and the **income statement** the two most important ones. These documents are treated as a whole, they must be clear enough so that the conveyed information is understandable and useful for the making of economic decisions, and they must pursue a faithful picture of the patrimony of the firm, its financial condition and the income it generates.
More specifically, the balance sheet is a snapshot of the firm's situation on one very specific day, usually the 31st of December. The balance sheet shows both sides of the same coin: on one side the investment taken on by a firm up until this date (classified in **assets**) and on the other side the financing committed by a firm in order to make such investment, up until this date as well (classified in **liabilities**).
The purpose of the income statement is to determine whether a firm has been successful or not during the whole year by means of calculating its **profit** (or, better said, its **net income**). In order to determine such outcome, many accounts play a key role: **sales**, **purchases**, **depreciation**, **interests**, **taxes**, etc.

It is clear then that the results arisen in these documents lead to financial implications which at the same time result in conclusions about a firm's performance. Like human beings, each firm is unique, therefore we will obtain different conclusions for each one. However, we can join similar firms together according to certain criteria and therefore study groups instead of individuals. This kind of firm study based on its accounting information is our starting point.

According to [6], **Financial Statement Analysis** consists of the application of analytical tools and techniques to financial statements in order to derive from them measurements, relationships and interpretations significant and useful enough for decision making, and again, within corporate performance.

The bottom line is that Financial Statement Analysis uses the financial statements elaborated according to Accounting in order to provide results to the measure of corporate performance.

The interested reader may follow and expand the whole discussion in [6] and [17].

2.2.2 Introduction to the Mathematics of Financial Operations

A **Financial Operation** is any exchange of monetary disposables among economic subjects in different periods of time.

There are two **Personal Elements**: the **Active Subject** possesses the disposables and chooses to give them during a certain amount of time. The **Passive Subject** receives the disposables and commits itself to their future devolution.
The **Objective Element** are the monetary disposables exchanged by the economic subjects.
In order for any exchange to take place, it must exist an agreement of mutual wills between the personal elements. This is the **Conventional Element**. The exchange is commanded by an equivalence between each subject's contributions. Usually, such an equivalence is formalised through a commercial contract.
A **Simple Financial Operation** is the exchange of a certain capital for another one, with different deferrals.

We denote (C,T) a certain monetary disposable, where C is the monetary amount and T is the deferral, measured from the origin of the operation, of the instant of this amount's availability. The set of all pairs of monetary disposables is

$$F = \{(C,T) | C \in \mathbb{R}^+ \cup \{0\}, T \in \mathbb{R}^+ \cup \{0\}\}.$$

We define an equivalence relation in F: two pairs of capitals are equivalent if and only if it exists an implicit functional relationship between the components of the exchanged capitals:

$$(C,T) \sim (C',T') \Leftrightarrow \exists\, f \mid f(C,C',T,T') = 0$$

The following are the properties that $\sim$ satisfies. The two last ones refer to a strictly financial meaning:

(i) *Reflexive*: $\forall (C,T) \in F \;\; (C,T) \sim (C,T) \Leftrightarrow \exists\, f \mid f(C,C,T,T) = 0$

(ii) *Symmetric*: $\forall (C,T),(C',T') \in F \;\; (C,T) \sim (C',T') \Rightarrow (C',T') \sim (C,T) \;\; \Leftrightarrow$
$\exists\, f \mid f(C,C',T,T') = 0 \Rightarrow \exists\, f' \mid f'(C',C,T',T) = 0$

(iii) *Transitive*: $\forall (C,T),(C',T'),(C'',T'') \in F \quad (C,T) \sim (C',T'),\;\; (C',T') \sim (C'',T'') \Rightarrow$
$(C,T) \sim (C'',T'') \;\; \Leftrightarrow$
$\exists\, f, f' \mid f(C,C',T,T') = 0,\;\; f'(C',C'',T',T'') = 0 \Rightarrow \exists\, f'' \mid f''(C,C'',T,T'') = 0$

(iv) *Interest Positivity*: Given two financial capitals of the same amount, it will be economically preferable that with the least deferral. It is also called **Principle of Future Capital's Underestimation**, and it can be formalised as:

$$(C,T) \sim (C',T'),\;\; C' = C + \Delta C,\;\; T' = T + \Delta T \Rightarrow \epsilon(\Delta C) = \epsilon(\Delta T),$$

where ϵ denotes the sign.

(v) *Homogeneity in regard to amount*: $(C,T) \sim (C',T') \Rightarrow \forall k \in \mathbb{R}^+ \cup \{0\}\; (kC,T) \sim (kC',T') \Leftrightarrow$
$\exists\, f \mid f(C,C',T,T') = 0 \Rightarrow \forall k \in \mathbb{R}^+ \cup \{0\} \;\; \exists\, f_k \mid f_k(kC,kC',T,T') = 0.$

Applying (v) for $C > 0$ and taking $k = \frac{1}{C}$, we obtain $f(C, C', T, T') = 0 \Rightarrow f_{\frac{1}{C}}(\frac{C}{C}, \frac{C'}{C}, T, T') = 0$.

If we can explicit $f_{\frac{1}{C}}(\frac{C}{C}, \frac{C'}{C}, T, T') = 0$, then $\exists f^* | f^*(T, T') = \frac{C'}{C} \Leftrightarrow C' = Cf^*(T, T')$. f^* is called the **financial factor**.

It is not our purpose to study neither the properties of f^* nor the different expressions that it can take. We relate to [2] for the interested reader.

However, it is interesting to us to determine f^* from an accounting perspective. We first need to introduce one of the accounting principles defined in the Spanish **Accounting General Plan**. This principle, which could be translated as the **Principle of the Ongoing Firm** (*Principio de Empresa en Funcionamiento*) states that it is tacitly assumed that a firm will continue to exist in the future, therefore all criteria applied do not consider at all the determination of the firm's value in order to transmit it or liquidate it.

In terms of financial operations, let (C, T) be the capital that a firm possesses at a certain instant of time and (C', T') the capital that a firm possesses at a later instant of time. Taking this principle into account, then the pair $\{(C, T), (C', T')\}$ must satisfy that $C' = C + Result$, where *Result* means a certain income generated by the firm during the period $T' - T$. Hence,

$$C' = C + Result \Leftrightarrow C' = C(1 + \tfrac{Result}{C}) \Leftrightarrow f^*(T, T') = 1 + \tfrac{Result}{C}.$$

From a financial perspective, the term $\frac{Result}{C}$ is generally called a **return**. It is a function of time, more specifically it is a function of $T' - T$. It is a way of indicating the performance obtained by the firm in the period $T' - T$. To our purposes we consider this period to be a year, so it is an **annual return**.

We are now going to take a closer look to it, giving more thorough definitions to these concepts.

2.2.3 Return

There can be many ways to define what we understand by the **return** on a firm. In line with the concept of corporate performance introduced in 1.1, we may also say that it is the source of the rewards required to compensate the stakeholders for the risks that they are assuming.
Likewise, there are many criteria by which return can be measured. Using Financial Statement Analysis, the **Return On Investment (ROI)**, which compares the income generated (*Result*) to the capital invested (C) is one of the most valid and most widely recognized measures. It allows us to compare it to alternative uses of capital and also to the return yielded by firms operating in similar situations. ROI relates income (reward) to the size of the capital that was needed to generate it. The investment of capital, which reflects on the volume of assets that a firm possesses, can always yield some return, and the riskier the investment is the higher the return required must be, in order to make it worthwhile.

So far, we have $ROI = \frac{Income}{Investment}$. What is important is to appropriately choose which *Income* and *Investment* use. There is no widely accepted convention on these terms, instead each pair simply results in a different way to determine ROI. For our purposes, we are going to choose two different measures of both income and investment, thus resulting in two different ways of determining ROI.

These four measures are four of the firm's accounts, and its values are taken in a specific moment in time. We refer to [6] or [17] in order to know about the nature of these accounts.

1. Using **Total Assets** as the investment:

When we compare the income generated to the total assets that a firm possesses at the start of the year, we determine a return that is specifically called **Return on Assets (ROA)**. In order to consider the income derived from strictly the operating activity of the firm, we take the **Earnings Before Interests and Taxes (EBIT)** as the income. Therefore, we have

$$ROI_{Total\ Assets} = ROA = \frac{EBIT_{31-12-t}}{Total\ Assets_{01-01-t}}\ .$$

2. Using **Total Equity** as the investment:

When we compare the income generated to the total equity of the firm at the start of the year, we determine a return that is specifically called **Return on Equity (ROA)**. In this case, we take the **Net Income** (EBIT - Interests - Taxes) as the income. Therefore, we have

$$ROI_{Total\ Equity} = ROE = \frac{Net\ Income_{31-12-t}}{Total\ Equity_{01-01-t}}\ .$$

In 2.2.2 we introduced an accounting principle that we must always bear in mind, the Principle of the Ongoing Firm. Right now we need to explain another one which is the **Principle of the Accrual** (*Principio del Devengo*). It states that the effects of all transactions or economic events in which the firm is involved will be registered whenever they occur their associated revenues or expenses, and therefore they will belong to the financial statements of the year that they refer to, regardless of the date of their associated payments or charges.
In short, this principle makes it clear that, in Accounting, revenues and expenses prevail over payments and charges.
We are taking this principle into consideration when determining ROI because the values are accounts taken from the **income statement** instead of the **statement of cash flows**. If we had used the latter then the principle involved is the **Principle of Cash** (*Principio de Caja*), in which payments and charges prevail over revenues and expenses.

The point here is that while all firms must comply with the income statement, the statement of cash flows is only mandatory to some of them. Reasonably enough, the Principle of Cash is not included in the list of the legal accounting principles. If we had used the Statement of Cash Flows, the basic term involved is $Cash\ Flow = Net\ Income + Depreciation$ and, for example, we could obtain a slightly different ROA ruled by the Principle of Cash. This approach is beyond the scope of our work, mainly because it involves dealing with data that are normally unavailable or hard to get.

In any case, for the interested reader we refer to [6], which gives a comprehensive account not only of this subsection but of the whole section.

2.2.4 Solvency

If return is one leg of our study, then the other leg is **solvency**. Again, the concept of solvency is sometimes ambiguous because it may be used in many different subjects. To our purposes, *we* define solvency as a firm's ability to comply with two different requirements: firstly, the ability to meet its financial commitments (**capability**), and secondly, the ability to do it in a specific moment in time (**punctuality**).
Bearing in mind that our study refers to annual returns, it is sensible to frame solvency within the same period of time and therefore talk about capability within a year and punctuality within a year.

Many references treat solvency and capability as synonymous and tend to forget punctuality or not distinguish it. We want to emphasize that in our work solvency depends on these two separate factors. Therefore, a firm may be solvent because it achieves capability or punctuality, but a firm is insolvent if it fails to achieve both capability and punctuality.

Another important thing is that with our definition of solvency we avoid talking about **liquidity**, which also appears in most references and may be also confusing. Normally, liquidity measures the degree to which a firm can meet its short-term obligations (one example would be according to [6]). Much as we could force to make this concept appear in our definition by considering short-term capability and long-term capability, we prefer to keep things simple and not to distinguish between short-term and long-term.

Because of this two-dimensional component, there are four types of firms according to solvency: firms with both high capability and high punctuality (firms that might have some sort of dominant position), firms with high capability but low punctuality (typically pharmacies), firms with high punctuality but low capability (typically banks), and firms with both low capability and punctuality (insolvent firms).

In order to measure both the capability and the punctuality of a firm using its accounting data, multiple ratios may be applied and therefore there is a potential danger of redundancy and overlapping. Instead of providing a long list of possible ratios, we prefer to make a selection coherent with our definitions. Contrary to the ratios of return, now we do not compare the end of the year to the start of it but all data refer to the end of the year. This is because the concept of return is associated with the yield of capital, implying an evolution in time, whereas the concept of solvency is associated with a firm's situation, implying a specific moment in time.

As in 2.2.3, the following measures involve some of the firm's accounts, and its values are taken in a specific moment in time, at the end of the year. We refer to [6] or [17] in order to know about the nature of these accounts.

1. Capability Ratios:

- $\frac{Current\ Assets - Inventories}{Total\ Assets}$

This ratio weighs some kinds of assets over the total volume. More specifically, we only take into consideration the most liquid assets, namely *Current Assets - Inventories* = ***Receivables*** *+* ***Cash***. Being the most liquid assets means that they are the ones most easily converted to cash.
The use of this ratio is justified in the idea that the more liquid assets a firm has the more able it will be to meet its financial obligations. We exclude the inventories because the volume of this account substantially varies among firms and also among sectors, and therefore it could be distorting in some cases.

- $\frac{Current\ Assets}{Total\ Liabilities}$

This ratio weighs the proportion of debt that a firm holds in comparison to those assets that are expected to be liquidated within a year. This ratio shows, in a specific instant, whether a firm can pay its debts thanks to liquidating its current assets. It would not be correct to consider the total volume of assets because, for instance, a firm would not normally liquidate its **Non Current Assets**, according to the Principle of the Ongoing Firm.

- $\frac{Total\ Equity}{Total\ Liabilities}$

Similarly to the previous ratio, this one describes the structure of the financing leg of the firm, being the investing leg the other one. The objective is to know whether a firm relies more on the capital given by its shareholders or on lent capital. The answer has great implications both for return and solvency itself, because there are two agents involved, shareholders and lenders, who demand different kinds of returns to their investment.

2. Punctuality Ratios:

- $Average\ Payment\ Period = \frac{Accounts\ Payable}{Purchases} \cdot 365$

By computing this ratio in terms of days, the result shows how long (how many days) does it take for the firm to pay its creditors. The volume that the firm owes to its creditors is in the account **Accounts Payable**.
Normally it is the longer the better, but the best way to interpret it is pairing it with the next ratio and determining the difference. Also, it is important to distinguish between Accounts Payable and debt; in terms of Accounting, Accounts Payable are only one portion of the total amount of debt.

- $Average\ Receivable\ Period = \frac{Accounts\ Receivable}{Sales} \cdot 365$

By computing this ratio in terms of days, the result shows how long (how many days) does it take for the firm's debtors to pay the firm. The volume that the firm owns to its debtors is in the account **Accounts Receivable**.
Normally it is the shorter the better, but the best way to interpret it is pairing it with the other ratio and determining the difference.

So, if $APP - ARP$ is positive, it means that a firm possesses enough cash which enables it to pay its creditors. If the difference is negative, the firm faces paying its creditors without having previously received cash from its debtors. Ideally, we would expect the difference to be equal to zero, meaning that there is no gap whatsoever regarding payments and receivables.

These two punctuality ratios only involve one part of the debt. The point is that they provide a good reference in order to determine what would happen with the other agents that the firm currently owes because these accounts very much represent the ordinary operating cycle of the firm.

2.2.5 Insolvency Risk: Connecting Return and Solvency

In this subsection we finally connect return and solvency, the two basic legs of Financial Statement Analysis. The main concept here is **risk**, and more specifically **insolvency risk**.

The firm's risk is the probability of loss, reflected in either repeatedly obtaining a negative income or experiencing a decrease in its assets' value (or an increase in its liabilities' value). If such is the outcome, then a firm might experience **financial distress** (in terms of risk it is common to name it **credit risk** or **default risk**, although some references like [17] treat them as different), and if it happens during a certain period of years there is a high chance that the firm ends up going **bankruptcy**.
Although this might be a correct approach to the issue, it does not match our purposes because there is no trace of insolvency risk and we want to only consider insolvency risk. Then, how do we redefine it?
The firm's risk may increase either because it increases the firm's inability to meet its long-term obligations or because it increases the firm's inability to meet its short-term obligations. In short, it increases the firm's inability to meet its obligations, involving both paying and complying with the dates. This is what we call insolvency risk, relating it to capability and punctuality.
It is important to remember that the inability to pay is strongly correlated with the inability to liquidate its currents assets.

When talking about return, we stated that the riskier the investment is the higher the return required must be, in order to make it worthwhile. Risk had not been introduced then. So, this means that there is a functional relation between risk and return.
More specifically and in terms of insolvency risk, what we are saying is that the higher the insolvency risk is the higher the return is expected to be and vice versa. In terms of ratios, the more positive a solvency ratio is, the more negative a return ratio should be, and vice versa.
This leads us to finally stating that return and solvency move in opposite directions. A firm would normally deal with the decision of whether to pursue large returns at the expense of potentially experiencing financial difficulties or the other way round.

For example, if a firm turns out to be solvent because its volume of assets exceeds its volume of liabilities and because it is punctual, this normally results in a large volume of cash readily available to utilise. The firm may be solvent but this is at the expense of offering a low return to its stakeholders because this cash accounts for capital that is not generating any kind of income.

Think of a government as a firm and that we want to invest in it. If we acquire government bonds, for instance, we expect them to have a low risk exposure. Another way to say this is that a government is *generally* solvent. Therefore, these kinds of products offer a low return. Contrary to this example we know that if we invest in a turbulent firm the risk is going to be higher, offering low prospects of solvency, but we are expecting a higher return in compensation.

There could arise the question of what about those firms that show good indicators of both return and solvency. Such an outcome can only be obtained if the firm finds itself to be in some sort of dominant position, where it is entitled to fix its own terms.

If a firm is not able to generate income (so it is facing low returns), and it finds it difficult to liquidate its assets in order to balance the situation, then the firm's risk may also increase, although this would depend on the amount of debt the firm owes.

The bottom line is that *in theory* there exists f such that return = f(solvency) = f(capability, punctuality) and, more specifically, that there generally is a **trade-off between return and solvency**.
Therefore, both return and solvency prove to be useful when assessing insolvency risk. This is why we detailed both return ratios and solvency ratios and we will certainly use both of them.

2.2.6 Consequences of Insolvency Risk

As mentioned previously, insolvency risk is usually associated with **financial distress**, **insolvency** (obviously), **failure**, **default** and **bankruptcy**. According to [4], although these terms are sometimes used interchangeably, they are distinctly different in their formal usage. The point is that there are actually not synonymous. We are not really interested in the exact differences but in the path that a Spanish firm might walk into if insolvency risk gradually increases. We specify Spanish because the legal procedure undertaken and the terms involved in it may vary across countries, although it is the same idea.

We recall that a firm gets to experience financial distress either due to the repeated obtaining of a negative income or due to a continuous decrease in its assets' value (or a continuous increase in its liabilities' value). If financial distress becomes regular, then we expect insolvency risk to progressively increase. Eventually, the firm can reach a stage where it is no longer able to proceed with its usual activity, therefore being compelled to start a legal procedure called ***Concurso de Acreedores***, which could be approximately translated as **Corporate Bankruptcy Reorganisation**. To use *compelled* is a bit misleading, since this procedure may start either because the firm itself agrees to make such a declaration or because some stakeholder demands it. An interesting remark here is that, in fact, a considerable amount of time can pass between the instant in which it is logical for the firm to make such a declaration and the instant in which it is effectively made. There are numerous reasons for this delay but the majority of them obey to a decision made by the own firm.

The objective of this procedure is to restructure the firm's accounts so that it can go back to a non-distressed state and to normal activity. More specifically, an agreement between the firm and its stakeholders can be reached involving either the exchange of debt for equity (debt's condonation) or the extension of the payment terms (swapping current liabilities for non current liabilities). In any case, it is a judicial process and it is treated as such.
Not all the reasons which lead a firm to this stage can be solved through this procedure, for example improving a firm's income. This obviously depends on multiple other factors and a firm has to deal with it differently, but it is sort of paradoxical.

Is it possible to reach such agreements? Well, for example, if a firm proves that despite its situation it has a high ROA, then there might be a chance to override the situation by making the suppliers become shareholders. But again, this depends on the judge's view and all the information available for his or her judgement.

If the legal process comes to an end without reaching an agreement, we call this situation bankruptcy or we say that the firm has **failed** or has gone bankrupt. In this case the process of liquidation starts: the firm must liquidate all its assets selling them and try to pay all its debts with the sell. Once the firm is liquidated the firm may be born again (start from scratch) or it ceases to exist.

Finally, we would like to give a figure, which can be consulted in the Spanish ***National Statistics Institute*** (**Instituto Nacional de Estadística (INE)**, `www.ine.es`), concerning the number of such legal procedures occurred in Spain. During the four-year period 2004-2007 there was registered a total of 3,069 cases, whereas during the seven-year period 2008-2014 that number sky-rocketed to 42,771 cases. In Spain, the first period is considered as of economic prosperity, while the second period is considered as of economic recession. This is just an empirical way to illustrate that normally, when a country does not face a tough economic situation, it is rather unlikely for a generic firm to inevitably reach this stage.

2.3 Statistical Theory

2.3.1 Elements of Multivariate Analysis

In order to introduce the Problem of Classification and some statistical methods that deal with it, we first need to present some basic elements of **Multivariate Analysis**.

According to [8], Multivariate Analysis is the area of **Statistics and Data Analysis** which studies, represents, analyses and interprets data consisting of observations of more than one variable from a sample of individuals.

We first look at the one-dimensional case (**Univariate Analysis**). Mathematically, a sample of $n > 1$ observations of a certain variable X (in our context that would be one ratio observed in n firms) is simply a set of n scalars $x_i \in \mathbb{R}$, $i = 1, ..., n$ which can be considered as a vector $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

We only have a finite number of observations of X, just a sample (we only have data from n specific firms). In terms of Probability, we only know n values of the random variable X.

The mean of $\boldsymbol{x}$ is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{1}^T \cdot \boldsymbol{x}$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector whose all components are 1 and T indicates the transposed vector.

The centred set of observations is $x_{i0} = x_i - \bar{x}$, $i = 1, ..., n$. In terms of $\boldsymbol{x}$,

$$\boldsymbol{x_0} = \begin{pmatrix} x_{10} \\ \vdots \\ x_{n0} \end{pmatrix} = \boldsymbol{x} - \mathbf{1} \cdot \bar{x}$$

We define the **total sum of squares** as $T = \sum_{i=1}^{n} x_{i0}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \|\boldsymbol{x_0}\|^2 = \|\boldsymbol{x}\|^2 - n\bar{x}^2$, where $\|\cdot\|$ denotes the Euclidean norm. Hence, the **variance** of $\boldsymbol{x}$ is $s_x^2 = \frac{1}{n}T$.

Now let us assume that the n observations are subdivided in $g > 1$ groups so that $n_\alpha = \#\{i \in \alpha\}$, $\alpha = 1, ..., g$ and $\sum_{\alpha=1}^{g} n_\alpha = n$ (in our context that would be $g = 2$ different groups, 'failed' firms and 'non-failed' firms). Without loss of generality, we may assume that the observations are ordered so that

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_g} \end{pmatrix},$$

$\boldsymbol{x_1} \in \mathbb{R}^{n_1}, ..., \boldsymbol{x_g} \in \mathbb{R}^{n_g}$. We calculate the g group centroids: $\bar{x}_\alpha = \frac{1}{n_\alpha}\sum_{i\in\alpha} x_i$, $\alpha = 1, ..., g$ and we center each group in its mean: $\boldsymbol{x_{\alpha 0}} = \boldsymbol{x_\alpha} - \mathbf{1_{n_\alpha}} \cdot \bar{x}_\alpha$, $\alpha = 1, ..., g$. Hence, $\bar{x} = \sum_{\alpha=1}^{g} \frac{n_\alpha}{n}\bar{x}_\alpha$.

We define the sum of squares within each group α as $W_\alpha = \|\boldsymbol{x_{\alpha 0}}\|^2 = \boldsymbol{x_{\alpha 0}}^T \cdot \boldsymbol{x_{\alpha 0}} = \sum_{i\in\alpha}(x_i - \bar{x}_\alpha)^2$.

Hence, the group variance is $s^2_{x_\alpha} = \frac{1}{n_\alpha} W_\alpha$, and the sum of these g sums is called the **within groups sum of squares**: $W = \sum_{\alpha=1}^{g} W_\alpha$.

Dividing it by n we obtain the so-called **pooled within groups variance**: $S_{pl} = \frac{1}{n} W = \sum_{\alpha=1}^{g} \frac{n_\alpha}{n} s^2_{x_\alpha}$.

Now, let us consider the mean group vector

$$\boldsymbol{M} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_g \end{pmatrix},$$

which again can be centred with respect to $\bar{x}$:

$$\boldsymbol{M}_0 = \begin{pmatrix} \bar{x}_1 - \bar{x} \\ \vdots \\ \bar{x}_g - \bar{x} \end{pmatrix},$$

and calculate the so-called **between groups sum of squares**: $B = \sum_{\alpha=1}^{g} n_\alpha (\bar{x}_\alpha - \bar{x})^2$.

Proposition 2.3.1.1. *These sums of squares verify $T = W + B$.*

Proof: $x_i - \bar{x} = (x_i - \bar{x}_\alpha) + (\bar{x}_\alpha - \bar{x})$, $x_i \in \alpha$, $\alpha \in \{1, ..., g\} \Rightarrow (x_i - \bar{x})^2 = (x_i - \bar{x}_\alpha)^2 + (\bar{x}_\alpha - \bar{x})^2 + 2(x_i - \bar{x}_\alpha)(\bar{x}_\alpha - \bar{x}) \Rightarrow \sum_{i \in \alpha} (x_i - \bar{x})^2 = W_\alpha + n_\alpha (\bar{x}_\alpha - \bar{x})^2 \Rightarrow T = \sum_{\alpha=1}^{g} \sum_{i \in \alpha} (x_i - \bar{x})^2 = \sum_{\alpha=1}^{g} W_\alpha + \sum_{\alpha=1}^{g} n_\alpha (\bar{x}_\alpha - \bar{x})^2 = W + B$ □

Now we look at the multidimensional case. Let us now consider $n > 2$ p-dimensional vectors $\boldsymbol{x_i} \in \mathbb{R}^p$, $i = 1, ..., n$ which can be considered as a matrix $\boldsymbol{X} \in \mathcal{M}(\mathbb{R}^n, \mathbb{R}^p)$ usually called the **multivariate data matrix** (or simply the **data matrix**),

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1}^T \\ \vdots \\ \boldsymbol{x_n}^T \end{pmatrix} = \begin{pmatrix} x_{11} \cdots x_{1p} \\ \vdots \\ x_{n1} \cdots x_{np} \end{pmatrix}.$$

The mean vector of $\boldsymbol{X}$ is $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i} = \frac{1}{n} \mathbf{1}^T \cdot \boldsymbol{X}$. We centre each vector, $\boldsymbol{x_{i0}} = \boldsymbol{x_i} - \bar{\boldsymbol{x}}$, $i = 1, ..., n$. In terms of $\boldsymbol{X}$,

$$\boldsymbol{X_0} = \begin{pmatrix} \boldsymbol{x_{10}}^T \\ \vdots \\ \boldsymbol{x_{n0}}^T \end{pmatrix} = \boldsymbol{X} - \mathbf{1} \cdot \bar{\boldsymbol{x}} = \boldsymbol{J} \cdot \boldsymbol{X},$$

where $\boldsymbol{J} = \boldsymbol{Id_n} - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T$. $\boldsymbol{X_0}$ is usually called the **centred data matrix**.

The **total scatter matrix** is $\boldsymbol{T} = \sum_{i=1}^{n} \boldsymbol{x_{i0}} \cdot \boldsymbol{x_{i0}}^T = \boldsymbol{X_0}^T \cdot \boldsymbol{X_0} = \boldsymbol{X}^T \cdot \boldsymbol{J} \cdot \boldsymbol{X} \in \mathcal{M}(\mathbb{R}^p, \mathbb{R}^p)$, and the **covariance matrix** of $\boldsymbol{X}$ is $\boldsymbol{S} = \frac{1}{n} \boldsymbol{T}$. $\boldsymbol{T}$ (and therefore $\boldsymbol{S}$) is a symmetric matrix, as $\boldsymbol{T}^T = (\boldsymbol{X_0}^T \cdot \boldsymbol{X_0})^T = \boldsymbol{X_0}^T \cdot (\boldsymbol{X_0}^T)^T = \boldsymbol{X_0}^T \cdot \boldsymbol{X_0} = \boldsymbol{T}$.

Now, any linear combination of the centred sample $\boldsymbol{z} = \boldsymbol{X_0} \cdot \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^p \setminus \{(0, ..., 0)\}$, "transforms" n vectors into n scalars. What is more, the variance of these n scalars is

$$s_{\boldsymbol{z}}^2 = \frac{1}{n}\boldsymbol{z}^T \cdot \boldsymbol{z} = \frac{1}{n}\boldsymbol{b}^T \cdot \boldsymbol{X_0}^T \cdot \boldsymbol{X_0} \cdot \boldsymbol{b} = \boldsymbol{b}^T \cdot \boldsymbol{S} \cdot \boldsymbol{b} \geq 0.$$

If we assume that $\boldsymbol{S}$ is non-singular, then $\boldsymbol{b}^T \cdot \boldsymbol{S} \cdot \boldsymbol{b} > 0$.
A real squared matrix $M \in \mathcal{M}(\mathbb{R}^p, \mathbb{R}^p)$ is said to be **positive definite** if $v^T \cdot M \cdot v > 0$ for all $v \in \mathbb{R}^p \setminus \{(0, ..., 0)\}$. Therefore, in that case $\boldsymbol{S}$ is a positive definite matrix.

The **correlation matrix** of $\boldsymbol{X}$ is $\boldsymbol{R} = \boldsymbol{D_s^{-1}} \cdot \boldsymbol{S} \cdot \boldsymbol{D_s^{-1}}$, where $\boldsymbol{D_s} = diag(\sqrt{diag(\boldsymbol{S})})$, meaning that its diagonal elements are the square root of the diagonal elements of $\boldsymbol{S}$.

From the above definitions, $rg(\boldsymbol{X}) = rg(\boldsymbol{X_0})$ and $rg(\boldsymbol{S}) = rg(\boldsymbol{R})$. The proof that $rg(\boldsymbol{S}) = rg(\boldsymbol{X_0})$ may be found in [14].

Now, as with the one-dimensional case, let us assume that the n observations are subdivided in $g > 1$ groups so that the n rows of $\boldsymbol{X}$ are subdivided in $g > 1$ groups and $n_\alpha = \#\{i \in \alpha\}$, $\alpha = 1, ..., g$, $\sum\limits_{\alpha=1}^{g} n_\alpha = n$. Likewise, without loss of generality we may assume that the observations are ordered so that

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \vdots \\ \boldsymbol{X_g} \end{pmatrix},$$

$\boldsymbol{X_1} \in \mathcal{M}(\mathbb{R}^{n_1}, \mathbb{R}^p), ..., \boldsymbol{X_g} \in \mathcal{M}(\mathbb{R}^{n_g}, \mathbb{R}^p)$. We calculate the g group centroids: $\bar{\boldsymbol{x}}_{\boldsymbol{\alpha}} = \frac{1}{n_\alpha} \sum\limits_{i \in \alpha} \boldsymbol{x_i} = \frac{1}{n_\alpha}\boldsymbol{1_{n_\alpha}^T} \cdot \boldsymbol{X_\alpha}$, $\alpha = 1, ..., g$ and we center each group in its mean: $\boldsymbol{X_{\alpha 0}} = \boldsymbol{X_\alpha} - \boldsymbol{1_{n_\alpha}} \cdot \bar{\boldsymbol{x}}_{\boldsymbol{\alpha}} = \boldsymbol{J_\alpha} \cdot \boldsymbol{X_\alpha}$, $\alpha = 1, ..., g$, where $\boldsymbol{J_\alpha} = \boldsymbol{Id_\alpha} - \frac{1}{n_\alpha}\boldsymbol{1_{n_\alpha}} \cdot \boldsymbol{1_{n_\alpha}}^T$. Hence, $\bar{\boldsymbol{x}} = \sum\limits_{\alpha=1}^{g} \frac{n_\alpha}{n} \bar{\boldsymbol{x}}_{\boldsymbol{\alpha}}$.

We define the scatter matrix within each group α as $\boldsymbol{W_\alpha} = \boldsymbol{X_{\alpha 0}}^T \cdot \boldsymbol{X_{\alpha 0}} = \boldsymbol{X_\alpha}^T \cdot \boldsymbol{J_\alpha} \cdot \boldsymbol{X_\alpha}$.
The group covariance matrix is $\boldsymbol{S_\alpha} = \frac{1}{n_\alpha} W_\alpha$. We also assume that $\boldsymbol{S_\alpha}$ is non-singular, $\alpha = 1, ..., g$. Hence, the sum of these g matrices is called the **within groups scatter matrix**: $\boldsymbol{W} = \sum\limits_{\alpha=1}^{g} \boldsymbol{W_\alpha}$.
Dividing it by n we obtain the so-called **pooled within groups covariance matrix**: $\boldsymbol{S_{pl}} = \frac{1}{n}\boldsymbol{W} = \sum\limits_{\alpha=1}^{g} \frac{n_\alpha}{n} \boldsymbol{S_\alpha}$.
Now let us consider the matrix of group means

$$\boldsymbol{M} = \begin{pmatrix} \bar{\boldsymbol{x}}_{\boldsymbol{1}}^T \\ \vdots \\ \bar{\boldsymbol{x}}_{\boldsymbol{g}}^T \end{pmatrix},$$

which again can be centred with respect to $\bar{\boldsymbol{x}}$:

$$\boldsymbol{M}_0 = \begin{pmatrix} (\bar{\boldsymbol{x}}_{\boldsymbol{1}} - \bar{\boldsymbol{x}})^T \\ \vdots \\ (\bar{\boldsymbol{x}}_{\boldsymbol{g}} - \bar{\boldsymbol{x}})^T \end{pmatrix},$$

and calculate the so-called **between groups scatter matrix**: $\boldsymbol{B} = \sum\limits_{\alpha=1}^{g} n_\alpha (\bar{\boldsymbol{x}}_{\boldsymbol{\alpha}} - \bar{\boldsymbol{x}}) \cdot (\bar{\boldsymbol{x}}_{\boldsymbol{\alpha}} - \bar{\boldsymbol{x}})^T = \boldsymbol{M_0}^T \cdot \boldsymbol{D_n} \cdot \boldsymbol{M_0}$, where $\boldsymbol{D_n} = diag(n) = diag(n_1, ..., n_g)$.

Proposition 2.3.1.2. *These scatter matrices verify* $\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}$.

The interested reader may consult the proof, review and expand the whole subsection in [8], [14] or [16].

2.3.2 The Problem of Classification

The prediction of insolvency risk is, in terms of its mathematical approach, an example of a problem which in the statistical literature is called the **Problem of Classification**, framed within the field of **Pattern Recognition and Machine Learning**. The following notes are taken from [7] and [8], where they are explained in much greater detail.

Pattern Recognition and Machine Learning is concerned with the discovery of regularities in data and the use of these regularities to basically classify data into different categories, usually called groups or populations.

In this field it arises the so-called **Curse of Dimensionality**: having to deal with spaces of high dimensionality, comprising many input variables. Therefore, some statistical methods are also aimed to **dimensionality reduction**. About this issue, two examples would be **Principal Component Analysis (PCA)**, which we are not going to discuss, and **Linear Discriminant Analysis (LDA)**, which is discussed in the next subsection. On the one hand, PCA searches the direction which maximises the variance of the data within an homogeneous population. On the other hand, LDA searches the direction which maximises the group separation, therefore assuming that data are divided into different groups.

Before we continue, we must make it clear that, for convenience, in this subsection and in other subsections of 2.3, we do not consider the general case of g groups. Whenever it is not specified, it is assumed that $g = 2$.

To our purposes, what we specifically want is to classify a firm into two mutually excluding (complementary) groups: in insolvent firms or in solvent firms, based on certain available information of it.

So, let Ω_1, Ω_2 be any two populations and $X_1, ..., X_p$ p observed variables of them. We denote $\boldsymbol{x} = (x_1, ..., x_p)$ the observations of the variables of one individual ω. The Problem of Classification, also called **The Problem of Identification** or **Discriminant Analysis**, attempts to assign ω to one of the two populations. There are numerous real-life examples in which this objective makes sense apart from insolvency risk: deciding whether to grant a loan, determining whether a tumour is benign or malign, identifying the species to which a plant belongs,...

A **discriminant rule** is a criterion which allows us to assign ω knowing $(x_1, ..., x_p)$ through a **discriminant function** $D(x_1, ..., x_p)$. The rule for classification is:

If $D(x_1, ..., x_p) \geq 0$, we assign ω to Ω_1

Otherwise, we assign ω to Ω_2

This rule divides $\mathbb{R}^p$ in two regions

$$R_1 = \{\boldsymbol{x} \in \mathbb{R}^p \mid D(x) \geq 0\},\ R_2 = \{\boldsymbol{x} \in \mathbb{R}^p \mid D(x) < 0\}.$$

It is clear that we will fail if we assign ω to a population which does not belong to. The probability of erroneous classification in terms of conditional probabilities is

$$ecp = P(R_2 \cap \Omega_1) + P(R_1 \cap \Omega_2) = P(R_2 \mid \Omega_1)P(\Omega_1) + P(R_1 \mid \Omega_2)P(\Omega_2).$$

In terms of densities,

$$ecp = P(\omega \in R_2, \Omega_1) + P(\omega \in R_1, \Omega_2) = \int_{R_2} P(\omega, \Omega_1)d\omega + \int_{R_1} P(\omega, \Omega_2)d\omega.$$

There are other approaches that do not use a discriminant function but these probabilities instead. They involve **Decision Theory** because it is a two-stage process: **inference and decision**. The idea is the following:
We are interested in the probabilities of the two populations conditioned to the observations, namely $P(\Omega_k \mid \omega)$, for each population Ω_k, $k = 1, 2$. We solve the inference problem of determining the population-conditional probabilities $P(\omega \mid \Omega_k)$ and the population probabilities $P(\Omega_k)$, and then using **Bayes' Theorem**, $P(\Omega_k \mid \omega) = \frac{P(\omega|\Omega_k)P(\Omega_k)}{P(\omega)} = \frac{P(\omega|\Omega_k)P(\Omega_k)}{\sum_{i=1}^{2} P(\omega|\Omega_i)P(\Omega_i)}$, $k = 1, 2$.
Once these probabilities are found, we decide the population of each new input ω based on, for example, minimising the probability of erroneous classification.

These probabilities are generally unknown and we are not going to determine them. Instead, when necessary we will have to make estimations of these probabilities using the information available, namely our samples.
As an example, when estimating the probability of erroneous classification, we use as an estimator the **misclassification rate**: it can be calculated using a **confusion matrix**, and it simply counts the number of mistakes made in the prediction, and then weighing these numbers with respect to the size of the sample.

2.3.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is sometimes also referred as **Fisher-LDA** or **Canonical Discriminant**.

Let us go back to Fisher's work, in which LDA was publicly introduced for the first time. In his 1936 article, quoted in 2.1.1, Fisher starts by saying: *when two or more populations have been measured in several characters* [...]*, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated.* In his context, there are four variables x_1, x_2, x_3, x_4 observed in two populations. So, LDA considers a linear discriminant function $X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$, the easiest possible option.
He goes on to say: [...] *the particular linear function which best discriminates the two species will be one for which the ratio D^2/S is greatest, by variation of the four coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ independently*, where D denotes the difference between the means of X and S denotes a linear combination of the within groups sums of squares. At this stage it is still not clear what D and S exactly mean. We will soon clarify them.

Let us now use [5] and [16] to formalise these latest concepts and the rest of the subsection.

In order to determine the desired linear discriminant function for two groups we assume that the groups have the same population covariance matrix $\boldsymbol{\Sigma}$ but different mean population vectors $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$.
So, using the same notation as in 2.3.1, let $\boldsymbol{X}$ be a sample of size $n > 2$ divided in two groups of size $n_1 > 1$ and $n_2 > 1$, $n_1 + n_2 = n$, so that

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{pmatrix},$$

$\boldsymbol{X_1} \in \mathcal{M}(\mathbb{R}^{n_1}, \mathbb{R}^p), \boldsymbol{X_2} \in \mathcal{M}(\mathbb{R}^{n_2}, \mathbb{R}^p)$. With the notation $\boldsymbol{x_{11}}, ..., \boldsymbol{x_{1n_1}}, \boldsymbol{x_{21}}, ..., \boldsymbol{x_{2n_2}}$, where $\boldsymbol{x_{ij}} \in \mathbb{R}^p$, $i = 1, 2$, $j = 1, ..., n_i$, it is

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_{11}}^T \\ \vdots \\ \boldsymbol{x_{1n_1}}^T \\ \boldsymbol{x_{21}}^T \\ \vdots \\ \boldsymbol{x_{2n_2}}^T \end{pmatrix}.$$

Let the difference in the sample group means be $\boldsymbol{d} = \bar{\boldsymbol{x}}_{\boldsymbol{1}} - \bar{\boldsymbol{x}}_{\boldsymbol{2}}$ and let the **sample (pooled within groups) covariance matrix** be

$$\widetilde{\boldsymbol{S_{pl}}} = \frac{1}{(n_1-1)+(n_2-1)}\boldsymbol{W} = \frac{1}{n-2}(\boldsymbol{W_1} + \boldsymbol{W_2}) = \frac{1}{n-2}((n_1-1)\boldsymbol{S_1} + (n_2-1)\boldsymbol{S_2}),$$

where $\boldsymbol{W}$ is the within groups scatter matrix *redefined* so that $\boldsymbol{S_\alpha} = \frac{1}{n_\alpha - 1}\boldsymbol{W_\alpha}$, $\alpha = 1, 2$. This factor change does not invalidate the previous results.

Although they look similar, $\widetilde{\boldsymbol{S_{pl}}} \neq \boldsymbol{S_{pl}} = \frac{1}{n}(n_1\boldsymbol{S_1} + n_2\boldsymbol{S_2})$.

Proposition 2.3.3.1. $\widetilde{\boldsymbol{S_{pl}}}$ *is symmetric and positive definite.*

Proof: Two squared matrices A and B satisfy $(A+B)^T = A^T + B^T$, $(rA)^T = rA^T \; \forall r \in \mathbb{R}$ and $(A^T \cdot A)^T = A^T \cdot A$. In particular, $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are symmetric. Hence,

$\widetilde{\boldsymbol{S_{pl}}}^T = (\frac{1}{n-2}((n_1-1)\boldsymbol{S_1} + (n_2-1)\boldsymbol{S_2}))^T = \frac{n_1-1}{n-2}\boldsymbol{S_1}^T + \frac{n_2-1}{n-2}\boldsymbol{S_2}^T = \frac{n_1-1}{n-2}\boldsymbol{S_1} + \frac{n_2-1}{n-2}\boldsymbol{S_2} =$
$\frac{1}{n-2}((n_1-1)\boldsymbol{S_1} + (n_2-1)\boldsymbol{S_2} = \widetilde{\boldsymbol{S_{pl}}}$ and $\widetilde{\boldsymbol{S_{pl}}}$ is symmetric.

For $\alpha \in \{1,2\}$, $\boldsymbol{S_\alpha}$ is positive definite because the variance of $\boldsymbol{z} = \boldsymbol{X_{\alpha 0}} \cdot \boldsymbol{b} \in \mathbb{R}^n$,
$\boldsymbol{b} \in \mathbb{R}^p \setminus \{(0,...,0)\}$ is $s_{\boldsymbol{z}}^2 = \frac{1}{n_\alpha - 1}\boldsymbol{z}^T \cdot \boldsymbol{z} = \frac{1}{n_\alpha - 1}\boldsymbol{b}^T \cdot \boldsymbol{X_{\alpha 0}}^T \cdot \boldsymbol{X_{\alpha 0}} \cdot \boldsymbol{b} = \boldsymbol{b}^T \cdot \boldsymbol{S_\alpha} \cdot \boldsymbol{b} \geq 0$, and assuming that $\boldsymbol{S_\alpha}$ is non-singular, then $\boldsymbol{b}^T \cdot \boldsymbol{S_\alpha} \cdot \boldsymbol{b} > 0$.

Now, $\boldsymbol{b}^T \cdot \boldsymbol{S_\alpha} \cdot \boldsymbol{b} > 0 \Rightarrow \boldsymbol{b}^T \cdot r\boldsymbol{S_\alpha} \cdot \boldsymbol{b} = r\, \boldsymbol{b}^T \cdot \boldsymbol{S_\alpha} \cdot \boldsymbol{b} > 0 \; \forall r \in \mathbb{R}^+ \setminus \{0\} \Rightarrow \frac{n_\alpha - 1}{n-2}\boldsymbol{S_\alpha}$ is positive definite.

Finally, $\boldsymbol{b}^T \cdot (\frac{n_1-1}{n-2}\boldsymbol{S_1} + \frac{n_2-1}{n-2}\boldsymbol{S_2}) \cdot \boldsymbol{b} = (\boldsymbol{b}^T \cdot \frac{n_1-1}{n-2}\boldsymbol{S_1} \cdot \boldsymbol{b}) + (\boldsymbol{b}^T \cdot \frac{n_2-1}{n-2}\boldsymbol{S_2} \cdot \boldsymbol{b}) > 0$ and $\widetilde{\boldsymbol{S_{pl}}}$ is positive definite. □

Proposition 2.3.3.2. $\widetilde{\boldsymbol{S_{pl}}}$ *is non-singular.*

Proof: If $\widetilde{\boldsymbol{S_{pl}}}$ was singular, $det(\widetilde{\boldsymbol{S_{pl}}}) = 0 \Rightarrow \exists \boldsymbol{b}^* \in \mathbb{R}^p \setminus \{(0,...,0)\}$ satisfying $\widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b}^* = 0 \Rightarrow$
$\exists \boldsymbol{b}^* \in \mathbb{R}^p \setminus \{(0,...,0)\}$ satisfying $(\boldsymbol{b}^*)^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b}^* = 0$ and $\widetilde{\boldsymbol{S_{pl}}}$ would not be positive definite $\Rightarrow \widetilde{\boldsymbol{S_{pl}}}$ is non-singular. □

This proposition guarantees the existence of the inverse of $\widetilde{\boldsymbol{S_{pl}}}$.

Proposition 2.3.3.3. $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ *is also symmetric and positive definite.*

Proof: According to the definition of a matrix inverse, $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is the only matrix satisfying $\widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \widetilde{\boldsymbol{S_{pl}}} = \boldsymbol{Id_p}, \widetilde{\boldsymbol{S_{pl}}} \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} = \boldsymbol{Id_p}$. Transposing, we obtain that the matrix $(\widetilde{\boldsymbol{S_{pl}}}^{-1})^T$ satisfies $\widetilde{\boldsymbol{S_{pl}}} \cdot (\widetilde{\boldsymbol{S_{pl}}}^{-1})^T = \boldsymbol{Id_p}, (\widetilde{\boldsymbol{S_{pl}}}^{-1})^T \cdot \widetilde{\boldsymbol{S_{pl}}} = \boldsymbol{Id_p} \Rightarrow \widetilde{\boldsymbol{S_{pl}}}^{-1} = (\widetilde{\boldsymbol{S_{pl}}}^{-1})^T$ and $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is symmetric.

Also, $\forall \boldsymbol{b} \in \mathbb{R}^p \setminus \{(0,...,0)\}$ $\boldsymbol{b}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b} > 0$. Taking $\widetilde{\boldsymbol{b}} = \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{b} \neq (0,...,0)$, as $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is non-singular, we have $\widetilde{\boldsymbol{b}}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \widetilde{\boldsymbol{b}} = \boldsymbol{b}^T \cdot (\widetilde{\boldsymbol{S_{pl}}}^{-1})^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{b} = \boldsymbol{b}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{b} > 0$ and $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is positive definite. □

Proposition 2.3.3.4. $\widetilde{\boldsymbol{S_{pl}}}$ *is an unbiased estimator of the common population covariance matrix* $\boldsymbol{\Sigma}$.

Proof: In general, if θ is an unknown population parameter, an estimator $\widehat{\theta}$ of θ is said to be an unbiased estimator of θ if $E[\widehat{\theta}] = \theta$. In our case, we must proof that $E[\widetilde{\boldsymbol{S_{pl}}}] = \boldsymbol{\Sigma}$. Using the properties of sum and scalar product of the expectation,

$E[\widetilde{\boldsymbol{S_{pl}}}] = E[\frac{1}{n-2}((n_1-1)\boldsymbol{S_1} + (n_2-1)\boldsymbol{S_2}))] = \frac{n_1-1}{n-2}E[\boldsymbol{S_1}] + \frac{n_2-1}{n-2}E[\boldsymbol{S_2}]$.

For $\alpha \in \{1,2\}$, $\boldsymbol{S_\alpha}$ satisfies $E[\boldsymbol{S_\alpha}] = \boldsymbol{\Sigma_\alpha}$, where $\boldsymbol{\Sigma_\alpha}$ is the population covariance matrix of the α group. We refer to [16] for its proof.
Our hypothesis is that $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$. Hence,
$E[\widetilde{\boldsymbol{S_{pl}}}] = \frac{n_1-1}{n-2}\boldsymbol{\Sigma} + \frac{n_2-1}{n-2}\boldsymbol{\Sigma} = \frac{(n_1-1)+(n_2-1)}{n-2}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$. □

Now, let us consider any linear combination of the observed variables

$$\boldsymbol{z} = \boldsymbol{b}^T \cdot \boldsymbol{y} = (b_1,...,b_p) \cdot (y_1,...,y_p)^T = b_1y_1 + ... + b_py_p,$$

where $\boldsymbol{y}$ is any observation.

The difference in the sample means of this linear combination is $\bar{\boldsymbol{z}}_{\mathbf{1}} - \bar{\boldsymbol{z}}_{\mathbf{2}} = \boldsymbol{b}^T \cdot \boldsymbol{d}$, and the **sample variance** of $\mathbf{Z}$ is $\boldsymbol{S_Z} = \boldsymbol{b}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b}$, which is an estimator of the variance of $\boldsymbol{Z}$.
Now it is clear that, in terms of Fisher's notation, $X = \boldsymbol{z}$, $D = \boldsymbol{b}^T \cdot \boldsymbol{d}$ and $S = \boldsymbol{S_Z}$. Therefore, the objective is to maximise

$$\frac{(\boldsymbol{b}^T \cdot \boldsymbol{d})^2}{\boldsymbol{b}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b}}$$

with respect to $\boldsymbol{b}$.
A solution is $\boldsymbol{b} = \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{d}$ or any multiple of it. Thus, the maximising vector $\boldsymbol{b}$ is not unique but its direction. This is the vector perpendicular to the separation hyperplane, which in the case $g = 2$, is a line. Hence, $\boldsymbol{b}^T \cdot \boldsymbol{y}$ projects points $\boldsymbol{y}$ onto the line on which that ratio is maximised.
For this $\boldsymbol{b}$ we have $\bar{\boldsymbol{z}}_{\mathbf{1}} - \bar{\boldsymbol{z}}_{\mathbf{2}} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{d}$ and hence

$$\frac{(\boldsymbol{b}^T \cdot \boldsymbol{d})^2}{\boldsymbol{b}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{b}} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{d}.$$

The linear function $\boldsymbol{b}^T \cdot \boldsymbol{y} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{y}$ is the desired linear discriminant function. So, it is pretty straightforward to apply LDA, since we only require the sample group means $\bar{\boldsymbol{x}}_{\mathbf{1}}$, $\bar{\boldsymbol{x}}_{\mathbf{2}}$ and the sample (pooled within groups) covariance matrix $\widetilde{\boldsymbol{S_{pl}}}$.

Once it is determined, we can classify a future observation $\boldsymbol{y}$. To determine whether $\boldsymbol{y}$ is closer to $\bar{\boldsymbol{x}}_{\mathbf{1}}$ or $\bar{\boldsymbol{x}}_{\mathbf{2}}$, we see if $\boldsymbol{z}$ is closer to the transformed mean $\bar{\boldsymbol{z}}_{\mathbf{1}}$ or $\bar{\boldsymbol{z}}_{\mathbf{2}}$: we evaluate $\boldsymbol{z}$ for each observation $\boldsymbol{x_{1i}}$, $i = 1, ..., n_1$, and obtain $\boldsymbol{z_{11}}, ..., \boldsymbol{z_{1n_1}} \Rightarrow \bar{\boldsymbol{z}}_{\mathbf{1}} = \boldsymbol{b}^T \cdot \bar{\boldsymbol{y}}_{\mathbf{1}} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \bar{\boldsymbol{y}}_{\mathbf{1}}$. Similarly, $\bar{\boldsymbol{z}}_{\mathbf{2}} = \boldsymbol{b}^T \cdot \bar{\boldsymbol{y}}_{\mathbf{2}}$. Therefore, we assign $\boldsymbol{y}$ to the first group if $\boldsymbol{z}$ is closer to $\bar{\boldsymbol{z}}_{\mathbf{1}}$ than to $\bar{\boldsymbol{z}}_{\mathbf{2}}$ and we assign $\boldsymbol{y}$ to the second group if $\boldsymbol{z}$ is closer to $\bar{\boldsymbol{z}}_{\mathbf{2}}$ than to $\bar{\boldsymbol{z}}_{\mathbf{1}}$.
Let us prove that $\boldsymbol{z}$ is closer to $\bar{\boldsymbol{z}}_{\mathbf{1}}$ if $\boldsymbol{z} \geq \frac{1}{2}(\bar{\boldsymbol{z}}_{\mathbf{1}} + \bar{\boldsymbol{z}}_{\mathbf{2}})$:
Since $\bar{\boldsymbol{z}}_{\mathbf{1}} - \bar{\boldsymbol{z}}_{\mathbf{2}} = \boldsymbol{b}^T \cdot \boldsymbol{d} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{d} > 0$ because $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is positive definite, we have $\bar{\boldsymbol{z}}_{\mathbf{1}} > \bar{\boldsymbol{z}}_{\mathbf{2}}$.
Since $\frac{1}{2}(\bar{\boldsymbol{z}}_{\mathbf{1}} + \bar{\boldsymbol{z}}_{\mathbf{2}})$ is the midpoint, $\boldsymbol{z} \geq \frac{1}{2}(\bar{\boldsymbol{z}}_{\mathbf{1}} + \bar{\boldsymbol{z}}_{\mathbf{2}})$ implies that $\boldsymbol{z}$ is closer to $\bar{\boldsymbol{z}}_{\mathbf{1}}$.
To express the classification rule in terms of $\boldsymbol{y}$:
$\frac{1}{2}(\bar{\boldsymbol{z}}_{\mathbf{1}} + \bar{\boldsymbol{z}}_{\mathbf{2}}) = \frac{1}{2}\boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot (\bar{\boldsymbol{x}}_{\mathbf{1}} + \bar{\boldsymbol{x}}_{\mathbf{2}}) \Rightarrow$ we assign $\boldsymbol{y}$ to the first group if

$$\boldsymbol{b}^T \cdot \boldsymbol{y} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}} \cdot \boldsymbol{y} \geq \frac{1}{2}\boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot (\bar{\boldsymbol{x}}_{\mathbf{1}} + \bar{\boldsymbol{x}}_{\mathbf{2}}),$$

and we assign $\boldsymbol{y}$ to the second group if

$$\boldsymbol{b}^T \cdot \boldsymbol{y} = \boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot \boldsymbol{y} < \frac{1}{2}\boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot (\bar{\boldsymbol{x}}_{\mathbf{1}} + \bar{\boldsymbol{x}}_{\mathbf{2}}).$$

In terms of 2.3.2, $D(x_1, ..., x_p) = \boldsymbol{b}^T \cdot \boldsymbol{y} - \frac{1}{2}\boldsymbol{d}^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot (\bar{\boldsymbol{x}}_{\mathbf{1}} + \bar{\boldsymbol{x}}_{\mathbf{2}})$.

The procedure described consists in solving a problem of **relative eigenvectors and eigenvalues**. This can be seen more clearly if we consider the ratio which has to be maximised as the function $J(w) = \frac{w^T \cdot \boldsymbol{B} \cdot w}{w^T \cdot \boldsymbol{W} \cdot w}$, where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the scatter matrices defined in 3.2.1. Although the two ratios look apparently different, it can be proved that in the case of $g = 2$ groups they lead to the same results. For example, $\widetilde{\boldsymbol{S_{pl}}}$ and $\boldsymbol{W}$ are proportional, hence their inverse matrices are also proportional. This results in a different eigenvalue but the same eigenvector, the same direction after all.

We add this comment because it may possibly look more intuitive: the idea is to look for the greater separation *between* groups and the lower separation *within* groups.

2.3.4 Logistic Regression

We introduce this method using [8].

Contrary to LDA, **Logistic Regression** is a **regression model** for **Logistic Discriminant Analysis**. This model allows us to estimate the probability of an event which hinges on certain correlated variables.

First of all, let us introduce this analysis. Assume that an event E may happen or not in each of the individuals of a certain population (in our context, E is to face insolvency). We consider a binary variable y taking

$$y = 1 \text{ if } E \text{ happens, } y = 0 \text{ if } E \text{ does not happen.}$$

If the probability of E does not depend on other variables, if we denote $p := P(E)$, then the likelihood of an observation y is

$$L = p^y(1-p)^{1-y},$$

because $L = p$ if $y = 1$ and $L = 1 - p$ if $y = 0$.

If we have n independent observations $y_1, ..., y_n$, the likelihood is

$$L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i},$$

where $\sum_{i=1}^{n} y_i$ is the absolute frequency of E in the n observations. Using the **maximum likelihood estimation**, to estimate p we solve the **likelihood equation**

$$\frac{\partial}{\partial p} ln\, L = 0,$$

whose solution is $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the relative frequency of E in the n observations. The asymptotic distribution of $\widehat{p}$ is normal $N(p, \frac{p(1-p)}{n})$.

However, in our context the probability of E does depend on other variables. Suppose that p depends on certain variables $X_1, ..., X_p$ (the ratios in our case). If $\boldsymbol{x} = (x_1, ..., x_p)^T$ are the variable observations of a certain individual ω, the probability of E given $\boldsymbol{x}$ is $P(y = 1 \mid \boldsymbol{x}) =: P(\boldsymbol{x})$. The complementary probability is $P(y = 0 \mid \boldsymbol{x}) =: 1 - P(\boldsymbol{x})$.

The regression model fitted is

$$ln(\tfrac{P(\boldsymbol{x})}{1-P(\boldsymbol{x})}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = \beta_0 + \beta^T \cdot \boldsymbol{x},$$

where $\beta = (\beta_1, ..., \beta_p)^T$ is the regression parameters vector. This model is equivalent to assume the following probabilities in terms of $\boldsymbol{x}$:

$$P(\boldsymbol{x}) = \tfrac{e^{\beta_0+\beta^T\cdot\boldsymbol{x}}}{1+e^{\beta_0+\beta^T\cdot\boldsymbol{x}}},\ 1 - P(\boldsymbol{x}) = \tfrac{1}{1+e^{\beta_0+\beta^T\cdot\boldsymbol{x}}}$$

Given an individual ω and assuming that the parameters have been estimated, the logistic discriminant rule decides that ω possesses E if $P(\boldsymbol{x}) > \frac{1}{2}$ and it does not possess E if $P(\boldsymbol{x}) \leq \frac{1}{2}$.

In terms of the discriminant function $D(x_1, ..., x_p) = L_g(\boldsymbol{x}) = ln(\frac{P(\boldsymbol{x})}{1-P(\boldsymbol{x})})$, the logistic decision rule is:
If $L_g(\boldsymbol{x}) > 0$ then $y = 1$, if $L_g(\boldsymbol{x}) \leq 0$ then $y = 0$.

The likelihood of n observations $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$ is

$$L = \prod_{i=1}^{n} P(\boldsymbol{x_i})^{y_i}(1 - P(\boldsymbol{x_i}))^{1-y_i} \Rightarrow ln\, L = \sum_{i=1}^{n} y_i ln(P(\boldsymbol{x_i})) + (1 - y_i)ln(1 - P(\boldsymbol{x_i})).$$

We solve the equations

$$\frac{\partial}{\partial \beta_j} ln\, L = 0,\ j = 0, 1, ..., p$$

in order to find the maximum likelihood estimators of the parameters.
For $i = 1, ..., n$ we have $ln(P(\boldsymbol{x_i})) = \beta_0 + \beta^T \boldsymbol{x_i} - ln(1 + e^{\beta_0 + \beta^T \boldsymbol{x_i}})$, hence

$$\frac{\partial}{\partial \beta_0} ln(P(\boldsymbol{x_i})) = 1 - \frac{e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}}{1 + e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}} = 1 - P(\boldsymbol{x_i})$$

$$\frac{\partial}{\partial \beta_j} ln(P(\boldsymbol{x_i})) = x_{ij} - x_{ij} \frac{e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}}{1 + e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}} = x_{ij}(1 - P(\boldsymbol{x_i})),\ j = 1, ..., p.$$

Likewise, we have $ln(1 - P(\boldsymbol{x_i})) = -ln(1 + e^{\beta_0 + \beta^T \boldsymbol{x_i}})$, hence

$$\frac{\partial}{\partial \beta_0} ln(1 - P(\boldsymbol{x_i})) = -\frac{e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}}{1 + e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}} = -P(\boldsymbol{x_i})$$

$$\frac{\partial}{\partial \beta_j} ln(1 - P(\boldsymbol{x_i})) = -x_{ij} \frac{e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}}{1 + e^{\beta_0 + \beta^T \cdot \boldsymbol{x_i}}} = -x_{ij} P(\boldsymbol{x_i}),\ j = 1, ..., p.$$

We then obtain the likelihood equations to estimate the parameters,

$$\frac{\partial}{\partial \beta_0} ln\, L = \sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_0} ln(P(\boldsymbol{x_i})) + (1 - y_i) \frac{\partial}{\partial \beta_0} ln(1 - P(\boldsymbol{x_i})) = \sum_{i=1}^{n} (y_i - P(\boldsymbol{x_i})) = 0,$$

$$\frac{\partial}{\partial \beta_j} ln\, L = \sum_{i=1}^{n} y_i \frac{\partial}{\partial \beta_j} ln(P(\boldsymbol{x_i})) + (1 - y_i) \frac{\partial}{\partial \beta_j} ln(1 - P(\boldsymbol{x_i})) = \sum_{i=1}^{n} x_{ij}(y_i - P(\boldsymbol{x_i})) = 0,\ j = 1, ..., p.$$

Unfortunately, these equations cannot be solved explicitly and we must resort to iterative numerical procedures.

Finally, we must say that this model holds a problem in our context. It can be proved that the maximum likelihood estimators of the parameters do not exist if the population samples are completely separated, which is precisely the desired objective in the Problem of Classification. In that case, the parameters must be estimated differently.

2.3.5 k-Nearest Neighbours

The following notes are elaborated using [16].

The ***k*-Nearest Neighbours (KNN)** method is an example of a **non-parametric method**, contrary to LDA an Logistic Regression which are examples of **parametric methods**.
In terms of Probability, non-parametric methods try to estimate probabilities but making few assumptions about the form of the distributions involved. Let $\boldsymbol{y_1}, ..., \boldsymbol{y_n}$ be a sample of p observed variables, $\boldsymbol{y_i} \in \mathbb{R}^p, i = 1, ..., n$. We fix $k \in \{1, ..., n-1\}$.
For $i = 1, ..., n$ we calculate the distance from an observation $\boldsymbol{y_i}$ to all the other observations $\boldsymbol{y_j}, j = 1, ..., \widehat{i}, ..., n$ using the **Mahalanobis distance function**

$$(\boldsymbol{y_i} - \boldsymbol{y_j})^T \cdot \widetilde{\boldsymbol{S_{pl}}}^{-1} \cdot (\boldsymbol{y_i} - \boldsymbol{y_j}),\ j = 1, ..., \widehat{i}, ..., n,$$

where $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is the sample (pooled within groups) covariance matrix introduced in 2.3.3.
It is a distance because $\widetilde{\boldsymbol{S_{pl}}}^{-1}$ is symmetric and positive definite.

In order to classify $\boldsymbol{y_i}$ into one of the two groups, using the distance we take the k nearest observations to $\boldsymbol{y_i}$. We count the number of observations k_1 which belong to the first group and the number of observations $k_2 = k - k_1$ which belong to the second group. The classification rule is simple:
If $n = n_1 + n_2$ and $n_1 = n_2 \Rightarrow$ If $k_1 > k_2$ we assign $\boldsymbol{y_i}$ to the first group, otherwise we assign $\boldsymbol{y_i}$ to the second group.
If $n = n_1 + n_2$ and $n_1 \neq n_2 \Rightarrow$ If $\frac{k_1}{n_1} > \frac{k_2}{n_2}$ we assign $\boldsymbol{y_i}$ to the first group, otherwise we assign $\boldsymbol{y_i}$ to the second group.

A decision must be made as to the value of k. In practice, the most common election is to try several values of k and end up using the one with the lowest misclassification rate.

Although k-Nearest Neighbours method has a simple structure, it has some drawbacks. The most important one, specially when there is a high volume of data, is that the sample must be stored for every observation added. Also, there is the fact that k cannot change during the procedure.

2.3.6 Classification and Regression Trees

In this subsection we talk about a set of non-linear predictive models called **Prediction Trees**. We only deal with a specific type of Prediction Trees called **Classification and Regression Trees (CART)**. There are others like the **Iterative Dichotomiser 3 (ID3)** (and its descendants), **Automatic Interaction Detection (AID)** or **Chi-squared Automatic Interaction Detection (CHAID)**.

Let us first introduce some basic considerations about Prediction Trees.

A Prediction Tree consists of nodes and edges (or branches). It starts in the so-called root node. Each node represents an election between some alternatives which are in turn represented by the branches attached to the node. More specifically, the ramifications correspond to partitions of the set of observations. The partition is made according to a question about the value of a certain variable, and the criterion for the question sequence depends on what do we want to maximise.
The majority of Prediction Trees are binary, meaning that in each node there is an election between two alternatives, involving two predictors.
Each terminal node (or leaf) represents a value of the response variable.

Let us now describe the general procedure for CART.

This type of Prediction Trees consists of a **binary recursive partition**. CART accepts both quantitative and qualitative variables for predictor or response. The idea is to start from a root node which contains all sample cases and look for the binary distinction which gives us the most information possible about a group. We then take each of the resulting new nodes and repeat the same action, continuing the recursion until we reach the stopping criterion chosen.
One stopping criterion can be of node homogeneity (homogeneity maximisation), meaning that all the elements of a node have the same response value. Another possible stopping criterion can be that it stops when the number of elements in a terminal node is lower to a certain fixed value. Finally, a third stopping criterion can be in terms of the tree depth, meaning that it stops when the number of nodes divided by branches and leaves is greater than a certain fixed value.
Finally, once the tree is built, there is an optional second procedure of **pruning**, which means eliminating any irrelevant branches. This is evaluated using cross-validation, a general procedure beyond Prediction Trees which will be duly explained in 3.4.

In our context, Classification Trees try to predict a discrete category (a group) rather than a numerical value (being that the case of Regression Trees). The differences between these two types have to do with how the information is measured, what kind of predictions does the tree make, and how do we measure the prediction error.

The tree can make two types of predictions: a point prediction or a distributional prediction. The former simply predicts the group that sequence of nodes has led to, while the latter gives the estimate group probability.

We choose that the measurement of the prediction error is through the misclassification rate, as it is discussed in 2.3.2 and it is exemplified with our samples in 3.4.

The interesting part is the information measurement. We present the following two alternatives, although it is more common in CART to use the first one.

On the one hand, the **Gini Impurity Measure** shows how often a randomly chosen element from the step set would be incorrectly classified if it was classified according to the distribution of groups in the subset.
The Gini Impurity Measure of the relative frequency vector $\boldsymbol{f} = (f_1, ..., f_p)$, where f_i is the relative frequency of classifying into group i, $i = 1, ..., p$ is

$$G(\boldsymbol{f}) = \sum_{i=1}^{p} f_i(1 - f_i) = 1 - \sum_{i=1}^{p} f_i^2.$$

This measure reaches its minimum (zero) when all observations in the node fall into the same group. So, looking for the lowest value of this measure in each step encourages the formation of regions in which a high proportion of observations are assigned to one group.

On the other hand, the **entropy** is a basic concept within **Information Theory**, so we first give some notes about it using [7], [18] and [19], and referring to them for further details.

Information Theory, among many other uses, is one way to mathematically formalise ideas about uncertainty reduction and discrimination. This is why it is used alongside Probability Theory and Decision Theory.

Let X be a random variable. We ask how much information is received when we observe a specific value x of X. We may think of the amount of information as the 'degree of surprise' on learning the value $X = x$.

This information measure is a function $h(x)$ defined as $h(x) := -ln(P(x))$, if X is discrete, and $h(x) := -ln(f(x))$, if X is continuous with density f.
The lower probability an event has, the higher the information we receive.

We define the **entropy** of X as

$$H[X] := \sum_{x \in X} P(x)h(x) = -\sum_{x \in X} P(x)ln(P(x)),$$

if X is discrete, and

$$H[X] := \int_{Im(X)} f(x)h(x) = -\int_{Im(X)} f(x)ln(f(x)),$$

if X is continuous with density f, in which case it is called the **differential entropy**.
It indicates the value of X in terms of its average amount of information. Because $\lim_{p \to 0} p\, ln(p) = 0$, we shall take $P(x)ln(P(x)) = 0$ or $f(x)ln(x) = 0$ whenever x is such that $P(x) = 0$ or $f(x) = 0$.

Now assume that we have a discrete random vector (X, Y). If a value x of X is already known, then the additional information needed to specify the corresponding value of Y is given by $-ln(P(y \mid x))$, where $P(y \mid x) = \frac{P(y,x)}{P(x)}$ is the (conditional) probability of $X = x$ conditioned to $Y = y$.
The **conditional entropy of Y given $X = x$** is

$$H[Y \mid X = x] := -\sum_{y \in Y} P(y \mid x)ln(P(y \mid x)).$$

The average additional information needed to specify Y given X can be written as

$$H[Y \mid X] := \sum_{x \in X} P(x) H[Y \mid X = x] = - \sum_{x \in X} P(x) \left(\sum_{y \in Y} P(y \mid x) ln(P(y \mid x)) \right) =$$
$$= - \sum_{x \in X, y \in Y} P(y, x) ln(P(y \mid x)),$$

which is called the **conditional entropy of Y given X**.

The results are similar in the case of a continuous random vector (X, Y) with joint density $f_{(X,Y)}$ and marginal densities f_X, f_Y. If a value x of X is already known, then the additional information needed to specify the corresponding value of Y is given by $-ln(f(y \mid x))$, where

$$f(y \mid x) := \frac{f_{(X,Y)}(x,y)}{f_X(x)}$$

is the conditional density of Y with respect to $X = x$.
The conditional entropy of Y given $X = x$ in this case is

$$H[Y \mid X = x] := - \int_{Im(Y)} f(y \mid x) ln(f(y \mid x)) dy.$$

The average additional information needed to specify Y given X in this case can be written as

$$H[Y \mid X] := \int\limits_{Im(X)} f_x(x) H[Y \mid X = x] dx = - \int\limits_{Im(X)} f_x(x) \left(\int\limits_{Im(Y)} f(y \mid x) ln(f(y \mid x)) dy \right) dx =$$
$$= - \int\limits_{Im(X)} \left(\int\limits_{Im(Y)} f(y, x) ln(f(y \mid x)) dy \right) dx,$$

which is called the conditional entropy of Y given X.

Proposition 2.3.6.1. *A random vector (X, Y) satisfies $H[X, Y] = H[X] + H[Y \mid X]$, where $H[X, Y]$ is the **joint entropy** of (X, Y).*

Proof: We prove the equality in the discrete case.
The joint entropy of (X, Y) is defined as

$$H[X, Y] = - \sum_{x \in X, y \in Y} P(x, y) ln(P(x, y)),$$

so we must prove that

$$- \sum_{x \in X, y \in Y} P(x, y) ln(P(x, y)) = - \sum_{x \in X} P(x) ln(P(x)) - \sum_{x \in X, y \in Y} P(y, x) ln(P(y \mid x)).$$

$$\sum_{x \in X, y \in Y} P(y, x) ln(P(y \mid x)) - \sum_{x \in X, y \in Y} P(x, y) ln(P(x, y)) = \sum_{x \in X, y \in Y} P(y, x)(ln(P(y \mid x)) -$$
$$ln(P(x, y))) = \sum_{x \in X, y \in Y} P(y, x)(ln(P(y \mid x)) - ln(P(y, x))) = \sum_{x \in X, y \in Y} P(y, x)(ln(\tfrac{P(y,x)}{P(x)}) - ln(P(y, x))) =$$
$$- \sum_{x \in X, y \in Y} P(y, x) ln(P(x)) = - \sum_{x \in X} \left(\sum_{y \in Y} P(y, x) \right) ln(P(x)) = - \sum_{x \in X} P(x) ln(P(x)). \square$$

So, the information needed to describe X and Y is given by the sum of the information needed to describe X alone plus the additional information required to specify Y given X.

If (X, Y) is a random vector, the difference in entropies $H[Y] - H[Y \mid X = x] =: I[Y; X = x]$ is called the **realised information**, and it tells us how much our uncertainty about Y has changed thanks to observing $X = x$. Hence, the **expected information** (or the **mutual information**) of (X, Y) is

$$I[Y; X] := H[Y] - H[Y \mid X].$$

This is precisely what it can be used in every step of the classification tree in order to maximise the information gain. More specifically, in our context we consider a discrete random vector (Y, A) where Y is the response variable taking the g "values" $g_1, ..., g_p$, and A is the answer to some binary question about the predictors $X = (X_1, ..., X_p)$. Formally, $A = \mathbf{1}_{\mathcal{A}}(X)$ for some set $\mathcal{A}$. The idea is that, in each step, we measure how much do we learn about Y from knowing a certain A, namely $I[Y; A]$. Once the branches of a node have been determined, we repeat the procedure not with the whole sample but only with the observations of that node, according to recursive partition.

In order to determine this values we use the relative frequencies of the sample because the probabilities are unknown.

2.3.7 Neural Networks

The reference used in this subsection is [7].

Neural Networks have their origins in the attempts to find mathematical representations of information processing in biological systems. These networks are characterised by a high number of very simple units, like biological neurons, hence the name.
However, it has been used in a wide range of areas. We regard them as a set of efficient models for information processing and pattern recognition, therefore also potentially useful within the Problem of Classification.

The linear models for regression and classification are based on linear combinations of fixed basis functions $\phi_i(\boldsymbol{x})$ and take the general form

$$\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{b}) = f(\sum_{i=1}^{m} b_i \phi_i(\boldsymbol{x})),$$

where f is generally a non-linear function sometimes called an **activation function**. For example, it is the identity in the case of regression. Neural Networks extend the model by making $\phi_i(\boldsymbol{x})$ depend on parameters and then allow them to be adjusted, alongside b_i, during the so-called **training procedure**. This is why it is regarded as an **adaptive** method.

The model which we are interested in is the **feed-forward neural network**, also called the **multilayer perceptron**. Each unit with which this neural network is built is called **Rosenblatt's perceptron**, so we first explain it.

Rosenblatt's perceptron imitates the functioning of a biological neuron. Let $x_1, ..., x_p$ be p inputs which are multiplied by a weight $\boldsymbol{v} = (v_1, ..., v_p)$, obtaining the superposition

$$v_1 x_1 + ... + v_p x_p.$$

In terms of neurons, if this quantity exceeds a certain threshold c, the neuron is triggered and turns from a base state (-1) to an activity state (1).

Formally, let $(\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n)$ be n observations of p variables paired with their respective **target values**, $\boldsymbol{x_i} \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$, $i = 1, ..., n$. The perceptron is determined by the vectorial parameter $\boldsymbol{v} \in \mathbb{R}^p$ and the scalar $c \in \mathbb{R}$. The separation hyperplane is

$$\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^p \mid \boldsymbol{v}^T \cdot \boldsymbol{x} - c = 0\}.$$

For $i = 1, ..., n$, the prediction is

$$\hat{y}_i = \begin{cases} +1, & \text{if } \boldsymbol{v}^T \cdot \boldsymbol{x} - c \geq 0 \\ -1, & \text{if } \boldsymbol{v}^T \cdot \boldsymbol{x} - c < 0 \end{cases} = \epsilon(\boldsymbol{v}^T \cdot \boldsymbol{x} - c),$$

where ϵ denotes the sign. We can determine the corresponding error function based on the comparisons between y_i and $\hat{y}_i$.

One of its most important characteristics, which we will later see that it can also be applied to the multilayer perceptron, is its **learning capability**, the learning of the perceptron. This means that we can improve its prediction capability by minimising the error function.
A simple way of carrying this out is through an algorithm with an implemented iterative method such as the **Gradient Descent Method** or the **Newton-Raphson Method**.

The Gradient Descent Method is an iterative method which searches for a minimum of a function f. Given an approximate point z_i, the method calculates a better approximation $z_{i+1} := z_i - \eta f'(z_i)$, where η is the **learning speed**.

Newton-Raphson Method is just a variant of Gradient Descent Method. It calculates a solution of the equation $f'(z) = 0$ approximating f' through its tangent line. The tangent line passing through the point $(z_i, f'(z_i))$ to f' graphic crosses the x-axes in the point z_{i+1} which satisfies

$$0 - f'(z_i) = f''(z_i)(z_{i+1} - z_i).$$

Hence, $z_{i+1} = z_i - \frac{f'(z_i)}{f''(z_i)}$.

Let us illustrate Gradient Descent Method in the context of the perceptron. Assume that, for $i \in \{1, ..., n\}$, the chosen error function is

$$Error_i(\boldsymbol{v}) = [-y_i(\boldsymbol{v}^T \cdot \boldsymbol{x_i})]_+ = \begin{cases} -y_i(\boldsymbol{v}^T \cdot \boldsymbol{x_i}), & \text{if } -y_i(\boldsymbol{v}^T \cdot \boldsymbol{x_i}) \geq 0 \\ 0, & \text{if } -y_i(\boldsymbol{v}^T \cdot \boldsymbol{x_i}) < 0 \end{cases}.$$

Its derivative is $\frac{\partial Error_i(\boldsymbol{v})}{\partial \boldsymbol{v}} = -y_i \boldsymbol{x_i}$, hence $\boldsymbol{v_{k+1}} = \boldsymbol{v_k} + \eta y_i \boldsymbol{x_i}$ for a certain $k \in \mathbb{N}$. So, if we choose a learning speed $\eta \in \mathbb{R}^+$, in the k-step the algorithm would correct a misclassified point $(\boldsymbol{x_i}, y_i)$ by adjusting $\boldsymbol{v_{k+1}} = \boldsymbol{v_k} + \eta y_i \boldsymbol{x_i}$.

Let us now explain the multilayer perceptron. As previously stated, it is also called a feed-forward neural network. This is because these networks consist of various unit layers. The outputs of the first layer become the inputs of the second layer and so on. Strictly speaking this is called **function superposition**: the independent variable of a function is the dependent variable of the previous one.

First, we construct m linear combinations of the input variables $\boldsymbol{x} = (x_1, ..., x_p)$

$$a_i = \sum_{j=1}^{p} b_{ij}^{(1)} x_j \ + b_{i0}^{(1)},$$

where $i = 1, ..., m$ and (1) indicates that the corresponding parameters are in the first **layer** of the network. The parameters $b_{ij}^{(1)}$ are **weights** and the parameters $b_{i0}^{(1)}$ are **biases**. The quantities a_i are **activations**.

Each of them is then transformed using a differentiable non-linear activation function h to give

$$z_i = h(a_i),\ i = 1, ..., m,$$

the **hidden units**. Usual examples of the chosen h are the hyperbolic tangent or the logistic function $\sigma(a) = \frac{1}{1+e^{-a}}$.

These values are again linearly combined to give **output unit activations**

$$a_k = \sum_{i=1}^{m} b_{ki}^{(2)} z_i \ + b_{k0}^{(2)},\ k = 1, ..., n,$$

where n is the total number of outputs and (2) indicates that this transformation corresponds to the second layer of the network.

Finally, the output unit activations are transformed using another appropriate activation function $\tilde{h}$ to give a set of network outputs y_k. In this case, for standard regression problems the activation function is the identity, so that $y_k = a_k$, and for classification problems the activation function is the logistic function, so that $y_k = \sigma(a_k) = \frac{1}{1+e^{a_k}}$.

Hence, the overall network function for $k = 1, ..., n$ is

$$y_k(\boldsymbol{x}, \boldsymbol{b}) = \tilde{h}\left(\sum_{i=1}^{m} b_{ki}^{(2)} h(\sum_{j=1}^{p} b_{ij}^{(1)} x_j \ + b_{i0}^{(1)}) \ + b_{k0}^{(2)}\right),$$

where the set of all weight and bias parameters have been grouped together into a vector $\boldsymbol{b}$. In short, this Neural Network model is simply a non-linear function from a set of input variables $\{x_i\}$ to a set of output variables $\{y_k\}$ controlled by a vector $\boldsymbol{b}$ of adjustable parameters. The process of evaluating the function can be interpreted as a **forward propagation** of information through the network.

We have shown the multilayer perceptron for two layers. It can be easily generalised by processing multiple layers, each one consisting of a weighted linear combination

$$a_k = \sum_{i=1}^{m} b_{ki}^{(2)} z_i \ + b_{k0}^{(2)}, \ k = 1, ..., n$$

followed by an element-wise transformation using a certain non-linear activation function.

When comparing Rosenblatt's perceptron to the multilayer perceptron, we see that Rosenblatt's perceptron is the unit with which the multilayer perceptron is built. The multilayer perceptron is made of multiple units, and in each unit there is a vectorial function. While Rosenblatt's perceptron is restricted to linear predictions (linear functions involved), the multilayer perceptron obtains non-linear predictions (non-linear functions involved).

The same ideas of the learning of Rosenblatt's perceptron can be applied to the multilayer perceptron. More specifically, in the Problem of Classification in two groups, the procedure above described can be applied.

3 Practical Framework

3.1 SABI Database

SABI stands for ***Sistema de Análisis de Balances Ibéricos***, which could be translated as **Analysis System of Iberian Balances**. It is a database offered by *Bureau Van Dijk*, a globally-specialised firm in gathering and analysing data. We are entitled to a basic subscription of the database as members of the *Universitat de Barcelona (UB)*.

SABI contains financial and accounting information over 2.000.000 Spanish firms and over 500.000 Portuguese firms, although we will not need the Portuguese ones.

Once a firm is selected, it contains general details about it (see Figure 3.1.1) and accounting data (see Figures 3.1.2 and 3.1.3), for instance we find there the balance sheet and the income statement for a certain period. In some cases there may be available other financial statements like the statement of cash flows, but we will not need it either.

Our task is to build Sample 1 gathering the necessary data from SABI. For every firm belonging to the sample, we export all the data needed to a spreadsheet so that we can treat it more conveniently.

The following three figures illustrate three snapshots of SABI.

We refer to [10] in order to know more about how to use this database.

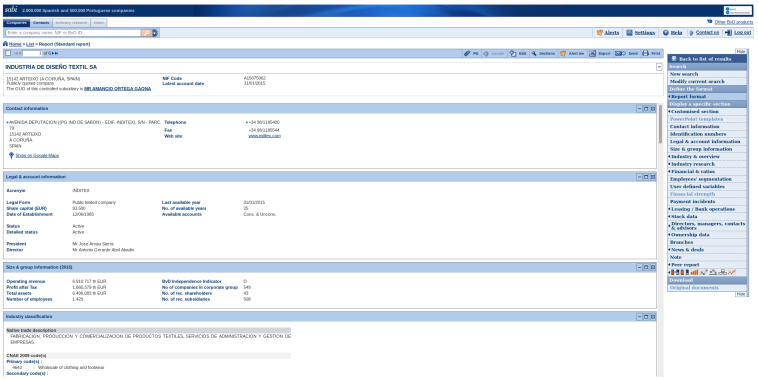

Figure 3.1.1: General details of INDUSTRIA DE DISEÑO TEXTIL SA

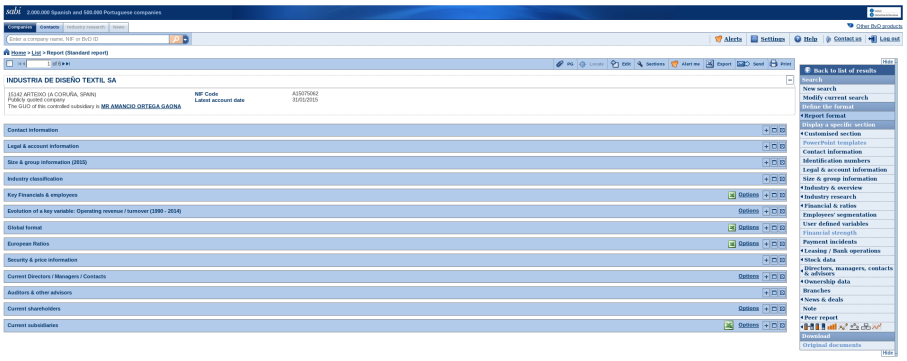

Figure 3.1.2: Overview of the available information of INDUSTRIA DE DISEÑO TEXTIL SA

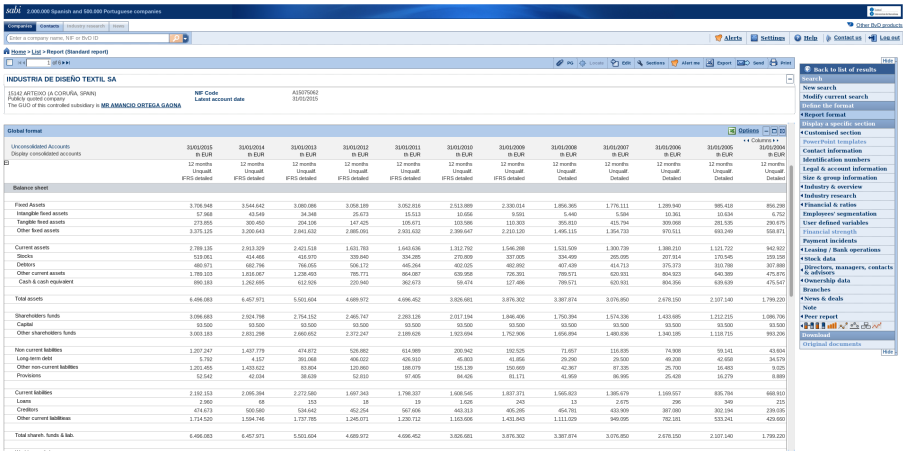

Figure 3.1.3: Balance sheets of INDUSTRIA DE DISEÑO TEXTIL SA

3.2 Sample 1 Details and Selected Predictors

Sample 1 contains two types (or groups) of firms. In line with the theoretical framework, one group consists of firms that have started that legal procedure so-called Corporate Bankruptcy Reorganisation, and the other group have not. Just for our convenience, we will refer to the first group as 'failed' firms and to the second group as 'non-failed' firms. Strictly speaking this is incorrect, as these terms are not synonymous.

In order to build this sample we use a **Case-Control Procedure**, meaning that we first add a failed-type firm and then we paired it with one (or more than one) non-failed-type firm. The pairing is done by selecting non-failed-type firms that had not started the legal procedure the same year than the failed-type firm. We decide to use a 1:3 proportion, so we select three non-failed-type firms for each failed-type firm.
The using of this procedure and this proportion is justified in the fact that it is reasonable to assume that only a quite small number of Spanish (and in general) firms end up starting the legal procedure. It would not be reasonable, for example, that the sample had the same number of the two types of firms. That would be somehow implying that the probability of insolvency is 0.5, which the experience tells us that it is certainly not the case.
Consequently, Case-Control Procedure is also used in epidemiology, for example.

More specifically, Sample 1 consists of 40 Spanish firms with the following characteristics:

- Of the 40 firms, 10 are failed-type firms and 30 are non-failed-type firms.

- They are Spanish firms legally constituted as either Incorporated (Inc., in Spanish *Sociedad Anónima (SA)*) or Limited (Ltd., in Spanish *Limitada (SL)*).

- They belong to sector "A(Agriculture, Forestry and Fishing).01: Crop and animal production, hunting and related service activities", according to **NACE Rev.2 Classification**. There is no specific reason behind the choosing of this sector, we could have chosen another sector. What it is important is that we must build a sample with firms belonging to the same sector, otherwise it would not be reasonable either because experience again tells us that the activity sector is a strong factor when it comes to compare similar firms.

- For each firm, we select data of only one specific year. In the case of the failed-type firms, it is the year in which the firm started the legal procedure. There are some firms which their year of start of the legal procedure is posterior to the last year of available data. In these cases, we decide that the year selected is the last year of available data. Once a failed-type firm is selected, the year selected for the non-failed-type firms paired with it is the same. The years selected belong to the period 2007-2014. The fact that the first possible year is 2007 is due to the Accounting General Plan: it was updated in 2007, so it is basically a legal reason behind the election. We cannot take data posterior to 2014 because at the time of the writing of this document the financial statements for 2015 are not finished yet.

Table 3.2.1 shows the list of the 10 failed-type firms alongside the 30 non-failed-type firms.

Sector	Failed Firm	Declaration of Failure Date (according to SABI)	Last Year with Available Data	Year (n) chosen as Declaration of Failure Year	Non-Failed Firms Associated with
01.	AGRICOLA VALLENIZA SL	26/05/2014	31/12/2012	2012	• 2007, SA • ABEREKIN SOCIEDAD ANONIMA • ABY SA
01.	AGRUPAEJIDO, SA	14/09/2015	31/08/2013	2013	• AFECAN SOCIEDAD ANONIMA • ACCIONES HORTICOLAS SA • ADOBAL SA
01.	CORTIJO MUDAPELO SA	03/01/2011	31/12/2012	2010	• ACAR SA • AGARCA SOCIEDAD ANONIMA • AGRALSA SA
01.	EIC AGRICOLA SA	30/10/2013	31/12/2009	2009	• AGRAR SYSTEMS SA • AGRARIA CASABLANCA SA • AGRARIA MANCHEGA HELLINERA SOCIEDAD ANONIMA
01.	AGROJUJUY SL.	28/04/2015	31/12/2012	2012	• 1329 PRUIT SL • 3M-AGROGAN SL • 7 GRUPO FERRELU SL
01.	AGROMAXIMO SL.	25/09/2013	31/12/2012	2012	• 919 MAS ROCAS SL • A CASTILLA NAZ SL • A D E RAMADERS SL
01.	ALFONSECA SL	24/09/2013	31/12/2012	2012	• A. CUTOLI, SL • A. FERRER SANCHO SOCIEDAD LIMITADA. • A. G. AREVALILLO E HIJOS S.L.
01.	ARROYO HONDO SL	07/09/2015	31/12/2013	2013	• A.P.G. COMERCIAL PALENCIA S.L. • A3 SISTEMAS HIDRAULICOS SL • ABAGRIOS SL
01.	CABANAS COMA RAMADERS SL	25/03/2014	31/12/2013	2013	• ABANCOR INTERNATIONAL TRADING SL • ABANICO DE PLATA SL • ABARANERA DE VERDURAS TOPI SL
01.	YEGUADA DE MILAGRO SA	04/09/2015	31/12/2013	2013	• ABAXTON NUTRICOM IBERICA SL • AGRAZUL SA • AGRICOLA ARENAS SA

Table 3.2.1: Firms of Sample 1

Each firm of the sample has seven observations so we can say that the sample is made of elements of $\mathbb{R}^7$. The seven observations are the seven ratios (or predictors) which we introduced in the theoretical framework and that we recall now:

$$R_1 = \frac{Current\ Assets - Inventories}{Total\ Assets}$$
$$R_2 = \frac{Current\ Assets}{Total\ Liabilities}$$
$$R_3 = \frac{Total\ Equity}{Total\ Liabilities}$$
$$R_4 = ROI_{Total\ Assets} = ROA = \frac{EBIT_{31-12-t}}{Total\ Assets_{01-01-t}}$$
$$R_5 = ROI_{Total\ Equity} = ROE = \frac{Net\ Income_{31-12-t}}{Total\ Equity_{01-01-t}}$$
$$R_6 = Average\ Receivable\ Period = \frac{Accounts\ Receivable}{Sales} \cdot 365$$
$$R_7 = Average\ Receivable\ Period = \frac{Accounts\ Receivable}{Sales} \cdot 365$$

Table 3.2.2 shows the resulting dataset (or sample observations) which can be viewed as a matrix $M \in (\mathbb{R}^{40}, \mathbb{R}^{7})$.

Firm	*Ratio1*	*Ratio2*	*Ratio3*	*Ratio4*	*Ratio5*	*Ratio6*	*Ratio7*	*Class*
AGRICOLA VALLENIZA SL	0.131	0.755	-0.090	-1.690	-0.160	151.145	67.658	Failed
2007, SA	0.680	5.976	6.248	0.048	0.035	73.226	13.329	Non-Failed
ABEREKIN SOCIEDAD ANONIMA	0.548	2.683	2.318	0.025	0.018	588.821	177.422	Non-Failed
ABY SA	0.057	0.550	2.126	0.002	0.011	54.674	202.840	Non-Failed
AGRUPAEJIDO, SA	0.165	0.434	-0.087	2.034	-0.039	5.211	28.108	Failed
AFECAN SOCIEDAD ANONIMA	0.480	1.431	1.211	-0.110	-0.055	277.388	757.001	Non-Failed
ACCIONES HORTICOLAS SA	0.227	1.274	3.288	0.058	0.063	43.498	50.344	Non-Failed
ADOBAL SA	0.327	1.563	2.498	0.058	0.067	62.778	217.184	Non-Failed
CORTIJO MUDAPELO SA	0.011	0.019	0.285	0.025	0.012	383.200	793.954	Failed
ACAR SA	0.082	0.275	1.575	0.006	-0.010	-100.261	195.046	Non-Failed
AGARCA SOCIEDAD ANONIMA	0.034	0.203	3.837	-0.957	-0.021	18.594	1186.188	Non-Failed
AGRALSA SA	0.375	2.351	2.963	-0.009	0.030	41.259	131.452	Non-Failed
EIC AGRICOLA SA	0.084	0.158	0.213	-0.062	-0.076	67.607	238.108	Failed
AGRAR SYSTEMS SA	0.384	0.969	0.197	0.076	0.034	171.484	87.800	Non-Failed
AGRARIA CASABLANCA SA	0.104	0.199	0.278	-0.070	0.012	116.566	50.931	Non-Failed
AGRARIA MANCHEGA HELLINERA SOCIEDAD ANONIMA	0.376	10.390	25.089	0.269	0.041	52.914	965.384	Non-Failed
AGROJUJUY SL.	0.143	0.431	1.383	-0.002	-0.015	110.987	230.462	Failed
1329 PRUIT SL	0.059	0.257	1.527	0.326	0.064	41.300	10.158	Non-Failed
3M-AGROGAN SL	0.100	1.394	0.451	0.002	0.020	119.509	17.474	Non-Failed
7 GRUPO FERRELU SL	0.462	2.421	3.311	0.031	0.008	3346.044	435.565	Non-Failed
AGROMAXIMO SL.	0.053	0.192	0.400	-0.020	0.020	3.176	62.090	Failed
919 MAS ROCAS SL	0.451	0.508	-0.110	-0.022	-0.131	66.363	39.442	Non-Failed
A CASTILLA NAZ SL	0.070	0.115	0.444	0.365	0.010	51.463	259.729	Non-Failed
A D E RAMADERS SL	0.114	0.957	1.396	1.077	0.089	40.176	22.825	Non-Failed
ALFONSECA SL	0.157	0.264	0.135	-0.107	-0.018	142.410	372.953	Failed
A. CUTOLI, SL	0.663	10.729	12.810	0.027	0.072	5.508	25.801	Non-Failed
A. FERRER SANCHO SOCIEDAD LIMITADA.	0.108	7.769	37.113	0.263	0.039	21.593	8.329	Non-Failed
A.G. AREVALILLO E HIJOS S.L.	0.046	0.539	0.112	0.000	0.029	256.420	35.921	Non-Failed
ARROYO HONDO SL	0.046	0.120	0.049	6.198	-0.053	139.870	240.820	Failed
A.P.G. COMERCIAL PALENCIA S.L.	0.380	0.785	0.261	-0.016	0.042	79.014	54.782	Non-Failed
A3 SISTEMAS HIDRAULICOS SL	0.288	0.774	-0.145	0.175	0.066	520.576	21.275	Non-Failed
ABAGRIOS SL	0.086	0.664	1.278	-0.160	0.020	119.705	44.060	Non-Failed
CABANAS COMA RAMADERS SL	0.116	0.150	-0.329	-1.629	-0.266	315.220	27.177	Failed
ABANCOR INTERNATIONAL TRADING SL	0.425	0.747	0.065	0.076	0.019	108.944	84.471	Non-Failed
ABANICO DE PLATA SL	0.135	0.733	-0.175	0.052	0.027	-317.293	72.464	Non-Failed
ABARANERA DE VERDURAS TOPI SL	0.288	0.451	0.561	-2.528	0.047	209.645	127.993	Non-Failed
YEGUADA DE MILAGRO SA	0.364	1.106	0.174	-0.051	-0.033	76.249	339.927	Failed
ABAXTON NUTRICOM IBERICA SL	0.554	1.008	0.095	0.138	0.051	90.531	129.870	Non-Failed
AGRAZUL SA	0.020	1.015	0.991	-0.212	-0.022	222.551	82.983	Non-Failed
AGRICOLAARENAS SA	0.183	0.934	0.038	-0.065	-0.020	63.665	7.140	Non-Failed

Table 3.2.2: Sample 1 observations

3.3 Sample 2 Details and Selected Predictors

Sample 2 (or Altman's Sample) is the sample used by E. I. Altman in his famous 1968 article. Its characteristics are described in [3], and we refer to it in order to fully understand how Altman carried out the process. The sample is composed of 66 American corporations with 33 firms in each group. The first group consists of **bankrupt** firms, namely manufacturers that filed a bankruptcy petition according to the **National Bankruptcy Act** during the period 1946-1965. The second group consists of data from the same period of firms which were still in existence in 1966.

As opposed to Sample 1, in this case bankrupt firms are not homogeneous because they do not belong to a specific sector. What Altman did to make up for this bias was to carefully select the non-bankrupt firms by stratifying firms by industry and by size, resulting in a paired group sample.

Each firm of the sample has five observations so we can say that the sample is made of elements of $\mathbb{R}^5$. The five observations are the following five ratios, selected by Altman:

$X_1 = \frac{Working\ Capital}{Total\ Assets}$

$X_2 = \frac{Retained\ Earnings}{Total\ Assets}$

$X_3 = \frac{Earnings\ Before\ Interest\ and\ Taxes}{Total\ Assets}$

$X_4 = \frac{Market\ Value\ of\ Equity}{Book\ Value\ of\ Total\ Debt}$

$X_5 = \frac{Sales}{Total\ Assets}$

Table 3.3.1 shows a fragment of the resulting dataset (or sample observations) which can be viewed, as a whole, as a matrix $M \in (\mathbb{R}^{66}, \mathbb{R}^5)$.

ID	Class	X1	X2	X3	X4	X5
B01	B	36.7	-62.8	-89.5	54.1	1.7
B02	B	24.0	3.3	-3.5	20.9	1.1
B03	B	-61.6	-120.8	-103.2	24.7	2.5
B04	B	-1.0	-18.1	-28.8	36.2	1.1
B05	B	18.9	-3.8	-50.6	26.4	0.9
B06	B	-57.2	-61.2	-56.2	11.0	1.7
B07	B	3.0	-20.3	-17.4	8.0	1.0
B08	B	-5.1	-194.5	-25.8	6.5	0.5
B09	B	17.9	20.8	-4.3	22.6	1.0
B10	B	5.4	-106.1	-22.9	23.8	1.5
B11	B	23.0	-39.4	-35.7	69.1	1.2
B12	B	-67.6	-164.1	-17.7	8.7	1.3
B13	B	-185.1	-308.9	-65.8	35.7	0.8
B14	B	13.5	7.2	-22.6	96.1	2.0
B15	B	-5.7	-118.3	-34.2	21.7	1.5
B16	B	72.4	-185.9	-280.0	12.5	6.7

Table 3.3.1: Fragment of Sample 2 observations

3.4 Results Obtained

In order to apply the previously defined statistical methods on the two samples we use **R**, a free statistical software. Since we only show here the final results, the R scripts used are attached in Appendix A to understand how the whole process was carried out. For example, for each sample the first script (LDA script) contains basic descriptive statistics prior to computing the method itself.

Before showing the results obtained for every method, we need to explain a couple of statistical concepts that are required to understand how can R produce the results. More specifically, in some methods, R allows us to use two types of step-wise procedures called **Leave-one-out (LOO)** and **Akaike Information Criterion (AIC)**. Let us briefly explain them.

The portion of a dataset (it may be the whole of it) utilised to fit a model is called the **training set**. According to [7], the performance on the training set may be in some cases not a good indicator of the predicted performance on unused data due to the problem of over-fitting, which arises when too many parameters are estimated.
If available data are plentiful, it is possible to only use some of them to fit the model and then test its prediction capability (and compare it to other models) using a so-called **validation set**. In the case that data are scarce, even if we divide our sample into a training set and a validation set, it may give a relatively noisy estimation of the predictive performance (due to over-fitting or other reasons), therefore resulting in a higher probability of erroneous classification.

In order to improve this estimation, a **cross-validation** procedure can be used: if S denotes a certain number of parts in which our data can be divided, it allows a proportion $\frac{S-1}{S}$ of it to be used for training while making use of all of the data to assess performance.

The LOO procedure considers $S = N$, where N is the size of the sample. LOO is a practical way to cross-validate when the sample is small. Therefore, when testing our samples using LDA, Logistic Regression and KNN in R, we use LOO technique so that in every step the training set is of size $N - 1$ and the prediction is applied to the remaining observation.

The "plug-in" procedure, which uses the same training set, is another way to estimate, but contrary to cross-validation, it provides a biased estimator.

According to [7], another procedure to specially avoid over-fitting is actually a set of criteria called **information criteria**. These criteria attempt to strike a balance between the maximum likelihood estimation and the number of parameters. They penalise the addition of parameters to increase the maximum likelihood estimation. One of these criteria is the AIC, defined in [1] (see the reference for more precision and deeper understanding) as

$$AIC = -2ln(\text{maximum likelihood}) + 2k,$$

where k is the number of estimated parameters. So, the larger the AIC is, the worse the model is deemed.
When testing our samples using Logistic Regression in R, we use AIC criterion so that in every step the objective is to improve the previous AIC calculated by varying the number of estimated parameters. We do not use this criterion in the LDA because it is not implemented in R, although it could be done manually.

Let us now present the results obtained in terms of 2.3 notation. In all the results concerning Sample 1, we assume that the first group is the failed-type firms group and the second group is the non-failed-type firms group (and similarly for Sample 2). We recall that we only use Sample 2 (Altman's Sample) in LDA, in order to see whether we obtain the same results.

LDA Results:

Applying the procedure described in 2.3.3 to Sample 1, the coefficients vector has the following components:
$b_1 = -5.368213, b_2 = 0.2991109, b_3 = -0.1324579, b_4 = 0.7307353, b_5 = -30.37555,$
$b_6 = -0.0002449891, b_7 = 0.001235236.$
Dividing the vector by b_7, for example, we obtain:
$b_1 = -4345.902, b_2 = 242.1489, b_3 = -107.2329, b_4 = 591.5756, b_5 = -24590.89,$
$b_6 = -0.1983339, b_7 = 1.$
Logically, we obtain the same results if we directly use the 'lda' function in R.
Hence, the linear discriminant function is $\boldsymbol{b}^T \cdot \boldsymbol{y} = (b_1, ..., b_7) \cdot (y_1, ..., y_7)^T = -4345.902y_1 + 242.1489y_2 - 107.2329y_3 + 591.5756y_4 - 24590.89y_5 - 0.1983339y_6 + y_7.$

The confusion matrix obtained (using LOO) is

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} 10 & 0 \\ 10 & 20 \end{pmatrix},$$

where c_{11} is the number of failed-type firms correctly classified as failed-type, c_{12} is the number of failed-type firms incorrectly classified as non-failed-type, c_{21} is the number of non-failed-type firms incorrectly classified as failed-type, and c_{22} is the number of non-failed-type firms correctly classified as non-failed-type. Hence, $c_{11} + c_{12} + c_{21} + c_{22} = n_1 + n_2 = n.$

The misclassification rate is $\widehat{ecp}_{LDA}^{Sample\,1} = 1 - \frac{c_{11}+c_{22}}{n} = 1 - \frac{10+20}{40} = \frac{10}{40} = 0.25.$

Applying the procedure described in 2.3.3 to Sample 2, the coefficients obtained are:
$b_1 = -0.015380337, b_2 = -0.018263460, b_3 = -0.041833488, b_4 = -0.007699163,$
$b_5 = -1.254215201.$
If we divide the coefficients by b_5 and round them to three decimals, we obtain:
$b_1 = 0.012, b_2 = 0.015, b_3 = 0.033, b_4 = 0.006, b_5 = 1.000,$ which are the same ones appearing in Altman's 1968 article.
Logically, we obtain the same results if we directly use the 'lda' function in R.
Hence, the linear discriminant function is $\boldsymbol{b}^T \cdot \boldsymbol{y} = (b_1, ..., b_5) \cdot (y_1, ..., y_5)^T = 0.012y_1 + 0.015y_2 + 0.033y_3 + 0.006y_4 + y_5.$

The confusion matrix obtained (using LOO) in this case is

$$\begin{pmatrix} 17 & 16 \\ 16 & 17 \end{pmatrix}.$$

The misclassification rate is $\widehat{ecp}_{LDA}^{Sample\,2} = 1 - \frac{17+17}{66} = \frac{16}{33} = 0.\,\widehat{48}.$

Logistic Regression Results:

Using the 'glm' function in R, the confusion matrix (without using LOO) is

$$\begin{pmatrix} 7 & 3 \\ 1 & 29 \end{pmatrix}.$$

The misclassification rate is $\widehat{ecp}_{Log.\,Reg.}^{Sample\,1} = 1 - \frac{7+29}{40} = \frac{1}{10} = 0.1.$

Using the 'glm' function in R, the confusion matrix (using LOO) is

$$\begin{pmatrix} 6 & 4 \\ 3 & 27 \end{pmatrix}.$$

The misclassification rate is $\widehat{ecp}^{Sample\,1}_{Log.\,Reg.} = 1 - \frac{6+27}{40} = \frac{7}{40} = 0.175$.

We recall that the maximum likelihood estimators of the parameters do not exist if the population samples are completely separated, and therefore they must be estimated differently. This is why, depending on the data provided, R could not apply well the method using the 'glm' function and give warnings because of it. This is what happens with Sample 1, we obtain results but we cannot entirely trust them. Instead, we need to use other functions which solve this issue, for instance the 'glmnet' function, contained in a specific R package with the same name.

KNN Results:

For $k \in \{1, 2, 3\}$, using the 'knn.cv' function in R, the confusion matrices and the misclassification rates obtained (using LOO) are the following:

· $k = 1$:

$$\begin{pmatrix} 5 & 5 \\ 4 & 26 \end{pmatrix}, \quad \widehat{ecp}^{Sample\,1}_{KNN} = 1 - \frac{5+26}{40} = \frac{9}{40} = 0.225$$

· $k = 2$:

$$\begin{pmatrix} 4 & 6 \\ 5 & 25 \end{pmatrix}, \quad \widehat{ecp}^{Sample\,1}_{KNN} = 1 - \frac{4+25}{40} = \frac{11}{40} = 0.275$$

· $k = 3$:

$$\begin{pmatrix} 4 & 6 \\ 2 & 28 \end{pmatrix}, \quad \widehat{ecp}^{Sample\,1}_{KNN} = 1 - \frac{4+28}{40} = \frac{1}{5} = 0.2$$

CART Results:

Using the 'rpart' function in R, Figure 3.4.1 shows the tree obtained, using Gini Impurity Measure.

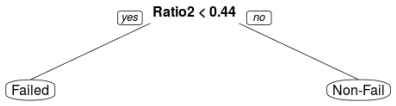

Figure 3.4.1: Tree obtained, using Gini Impurity Measure, for Sample 1

The confusion matrix and misclassification rate are

$$\begin{pmatrix} 8 & 2 \\ 5 & 25 \end{pmatrix}, \quad \widehat{ecp}^{Sample\,1}_{CART} = 1 - \frac{8+25}{40} = \frac{7}{40} = 0.175$$

Using the 'rpart' function in R, Figure 3.4.2 shows the tree obtained, using entropy.

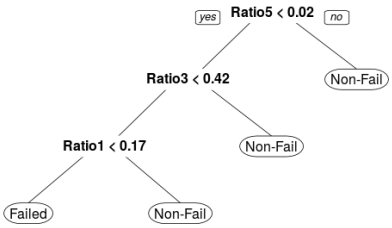

Figure 3.4.2: Tree obtained, using entropy, for Sample 1

The confusion matrix and misclassification rate obtained, using entropy, are

$$\begin{pmatrix} 8 & 2 \\ 1 & 29 \end{pmatrix}, \quad \widehat{ecp}_{CART}^{Sample\,1} = 1 - \tfrac{8+29}{40} = \tfrac{3}{40} = 0.075$$

Neural Networks Results:

Using the 'nnet' function in R, the confusion matrix and misclassification rate obtained, using two layers and 2 units in the hidden layer, are

$$\begin{pmatrix} 10 & 0 \\ 3 & 27 \end{pmatrix}, \quad \widehat{ecp}_{NN}^{Sample\,1} = 1 - \tfrac{10+27}{40} = \tfrac{3}{40} = 0.075$$

Using the 'nnet' function in R, the confusion matrix and misclassification rate obtained, using two layers and 3 units in the hidden layer, are

$$\begin{pmatrix} 8 & 2 \\ 7 & 23 \end{pmatrix}, \quad \widehat{ecp}_{NN}^{Sample\,1} = 1 - \tfrac{8+23}{40} = \tfrac{9}{40} = 0.225$$

4 Conclusions and Further Research

4.1 Conclusions

Our main study area is corporate performance, focusing on the characterisation and prediction of insolvency risk, which is a key point within it. By effectively predicting insolvency risk, two objectives of corporate performance can be achieved, which are the diagnosis of the situation of a firm in a specific moment of time and the answer to whether a firm is achieving its objectives, specifically putting the emphasis on the possibility of occurrence of financial distress and eventual failure.

The prediction of insolvency risk can be approached, for instance, from two different perspectives: using the so-called Financial Statement Analysis (basically Financial and Accounting Theory) or using Statistics. We have opted for using Financial and Accounting Theory only for its characterisation, while we have made use of Statistics for its prediction, presenting a series of appropriate statistical methods which prove to be reliable in the literature.

However, we have made it clear that there are connections between Accounting-Finance Theory and Statistical Theory, that it is quite pointless not to strike a balance between the two subjects. The most evident reason is that Accounting and Finance Theory turns out to be crucial in making the predictions, because it is through it that we justify the ratios selected as predictors. It is therefore most likely that the deeper is our understanding of it, the better is going to be the selection of ratios. In his 1968 article, E. I. Altman applied Linear Discriminant Analysis carefully selecting the ratios, so we have also given such an importance.
In short, Accounting and Finance Theory provides the conceptual background in which justify the results obtained with the methods.

We have based our analysis of insolvency risk in two linked legs, return and solvency. The main conclusion is that return and solvency normally move in opposite directions. This means that, *in general*, a firm earns high return rates at the expense of increasing its volume of debt, which in turn increases its insolvency risk.

We have formally presented five statistical methods with which test our samples and evaluate their prediction capability. The ordered misclassification rates obtained in 3.4, using Sample 1, are the following:

Method	**Specification**	**Misclassification Rate**
NN	2 layers and 2 units in hidden layer	7.5%
CART	Using entropy	7.5%
Log. Reg.	Without using LOO	10%
CART	Using Gini Impurity Measure	17.5%
Log. Reg.	Using LOO	17.5%
KNN	$k = 3$	20%
NN	2 layers and 3 units in hidden layer	22.5%
KNN	$k = 1$	22.5%
LDA	Using LOO	25%
KNN	$k = 2$	27.5%

So, we can state that both Neural Networks and Classification and Regression Trees have the greatest prediction capability among all the statistical methods which have been tested, taking into account the specifications and the fact that these results might depend on the sample used. More specifically, we have considered a multilayer perceptron with 2 layers and 2 units in the hidden layer, and a Classification Tree using entropy as information measure.

Looking at the order and the rates, we may affirm a couple of facts. In first place, that linear methods such as LDA, not only are they simple but they also provide acceptable enough results. In second place, that non-linear methods such as CART and NN are more complex but they provide better prediction results.
This confirms what we would have expected at the beginning, that there is a trade-off between complexity and prediction capability.

4.2 Further Research

When determining the specific characteristics of Sample 1, there are some potentially arguable aspects which could be analysed in further research.

We could ask themselves whether the seven ratios chosen as predictors are the best choice. Our purpose is to achieve the greatest prediction capability and that very much depends on the ratios used.
Our selection of ratios differs from that of Altman. Actually, we find numberless ratios in the literature. What we do is to select a small set of ratios justified on accounting and financial arguments so that they make sense. Although it is true that we have obtained good results in terms of prediction capability, we could determine the relative importance of each ratio and therefore decide if any change of variables would be useful. Or we could simply test other ratios trying to be coherent with the same arguments.

It could also be interesting to study more sectors or different groups of firms, and compare the results with those of Sample 1. What we do is to randomly choose firms from a specific sector for this sample, but we could ask ourselves to what extent the results obtained will be similar in other sectors. This involves choosing the most appropriate ratios in each case because it would be reasonable to assume that different sectors require different ratios, or that some ratios are better indicators only in some sectors.

Other considerations could be to examine the plausibility of the Case-Control relation, which in our case is 1:3, and the size of the sample, namely the number of firms selected.

Another option after obtaining the results, which we find in Altman's article, could be to repeat them but now using data from a different year, not one year prior to the declaration but two years prior, and then successively (three years prior,...), in order to assess how the prediction capability evolves if we change the period.

In short, we basically propose to extend and enrich the Practical Framework by adding both complements and alternatives to the way in which we proceeded.

On the other hand, the prediction capability also very much depends on the chosen method to make the predictions (or to classify the observations). So, it is reasonable to ask ourselves whether the five statistical methods chosen are the best choice.
Again, in 4.1 we corroborated the fact that these methods are reliable, but they are not the only ones which deal with the Problem of Classification. Other methods are quoted in 2.1.2, and through the references given we may find other examples.

In connection with the methods, there is the issue of how they are implemented in R. In theory, there should not be any difference between the method itself and the internal code implemented, but in some cases there is. For example, if we take LDA, the 'lda' function calculates the coefficients vector and then it normalises it (divides the vector by its Euclidean norm). Although the effect of normalising the vector does not alter the results (the direction stays the same), we should be aware of this additional step in the code.
Another case even more unsettling is when computing CART and NN, where 'rpart' and 'nnet' functions do not give enough information about how the methods are applied. In these two cases, we must admit that we have not looked up the details in the code, and therefore we only provide the final results without explaining the intermediate steps.

In conclusion, there is still a long way to go concerning the search for other appropriate statistical methods for the Problem of Classification and for prediction purposes, in general. Besides, whenever we test a method with an application (R in our case), we should not uncritically accept the given output but *ask for* the most possible information.

References

[1] Akaike, Hirotugu: A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control, Vol. AC-19, No. 6, December 1974.*

[2] Alegre Escolano, P. et al: Ejercicios resueltos de Matemática de las operaciones financieras, *Editorial AC. Madrid, 1989.*

[3] Altman, Edward I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance. Vol. XXIII, September 1968, No. 4 pp 589-609.*

[4] Altman, Edward I., Hotchkiss, Edith: Corporate Financial Distress and Bankruptcy, *Third Edition. John Wiley & Sons, Inc. Hoboken, New Jersey, 2006.*

[5] Anderson, T.W.: R.A. Fisher and Multivariate Analysis, *Statistical Science. 1996, Vol. 11, No. 1, pp 20-34.*

[6] Bernstein, Leopold A.: Financial Statement Analysis. Theory, Application, and Interpretation; *Fifth Edition. Richard D. Irwin, Inc. 1989.*

[7] Bishop, Christopher M.: Patter Recognition and Machine Learning, *Springer Science+Business Media, LLC. 2006.*

[8] Cuadras, Carles M.: Nuevos métodos de análisis multivariante, *CMC Editions. Barcelona, 2014.*
Available (Spanish version) at:
`http://www.ub.edu/stat/personal/cuadras/metodos.pdf`

[9] Fisher, R. Aylmer.: The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics, Volume 7 Issue 2: September 1936, pp 179-188.*

[10] Informa, S.A.: Manual del usuario SABI, *Julio de 2006.*

[11] López Iturriaga, Félix J., Pastor Sanz, Iván: Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks; *Expert Systems with Applications, No. 42 (2015), pp 2857-2869.*

[12] Olmeda, I., Fernández, E.: Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction, *Computational Economics, Vol. 10, pp 317-335. Kluwer Academic Publishers, 1997.*

[13] Platykanova, P.: El análisis económico-financiero: Estado del arte, *Revista de Contabilidad y Dirección. Vol. 2, año 2005, pp 95-120.*

[14] Rao, C. Radhakrishna: Linear Statistical Inference and Its Applications, *Second Edition. John Wiley & Sons, Inc. 1973.*

[15] Real Decreto 1514/2007, de 16 de noviembre, por el que se aprueba el
Plan General de Contabilidad; *Boletín Oficial del Estado, Suplemento del número 278, año CCCXLVII, martes 20 de noviembre de 2007.*

[16] Rencher, Alvin C.: Methods of Multivariate Analysis, *Second Edition. John Wiley & Sons, Inc. 2002.*

[17] Rojo Ramírez, Alfonso A.: Análisis Económico-Financiero de la Empresa. Un análisis desde los datos contables, *Garceta. Madrid, 2011.*

[18] Shalizi, Cosma: Finding Informative Features, *Notes of the course 36-350: Data Mining. 4 September, 2009.*
Available at:
http://www.stat.cmu.edu/∼cshalizi/350/lectures/05/lecture-05.pdf

[19] Shalizi, Cosma: Classification and Regression Trees, *Notes of the course 36-350: Data Mining. 6 November, 2009.*
Available at:
http://www.stat.cmu.edu/∼cshalizi/350/lectures/22/lecture-22.pdf

Appendices

A R Scripts

A.1 LDA Scripts

```
#
# Linear Discriminant Analysis (LDA)
#
# Sample: 40 firms (10 Failed and 30 Non-Failed) belonging to the sector
# 01. Crop and animal production, hunting and related service activities (according to
# NACE Rev. 2)
#
# Firms are either SA or SL, indistinctly
# Data from the 2007-2014 period (yearly)
#

# Read data table, stored in "TFG_Observations.txt" for example

firms<-read.table("TFG_Observations.txt",header=TRUE,row.names=1)

firms.failed<-firms[firms$Class=="Failed",-8]
firms.non.failed<-firms[firms$Class=="Non-Failed",-8]

# Table details and descriptive statistics

str(firms)

'data.frame':   40 obs. of  8 variables:
 $ Ratio1: num  0.131 0.68 0.548 0.057 0.165 ...
 $ Ratio2: num  0.755 5.976 2.683 0.55 0.434 ...
 $ Ratio3: num  -0.0897 6.2481 2.3175 2.1259 -0.0872 ...
 $ Ratio4: num  -1.6901 0.04801 0.02515 0.00211 2.03401 ...
 $ Ratio5: num  -0.1603 0.0354 0.0178 0.0111 -0.0389 ...
 $ Ratio6: num  151.14 73.23 588.82 54.67 5.21 ...
 $ Ratio7: num  67.7 13.3 177.4 202.8 28.1 ...
 $ Class : Factor w/ 2 levels "Failed","Non-Failed": 1 2 2 2 1 2 2 2 1 2 ...

names(firms)

[1] "Ratio1" "Ratio2" "Ratio3" "Ratio4" "Ratio5" "Ratio6" "Ratio7" "Class"

str(firms.failed)

'data.frame':   10 obs. of  7 variables:
 $ Ratio1: num  0.1314 0.1655 0.0105 0.0845 0.1434 ...
 $ Ratio2: num  0.755 0.434 0.019 0.158 0.431 ...
 $ Ratio3: num  -0.0897 -0.0872 0.2848 0.2133 1.3827 ...
 $ Ratio4: num  -1.6901 2.03401 0.02549 -0.0617 -0.00153 ...
 $ Ratio5: num  -0.1603 -0.0389 0.0124 -0.0757 -0.0153 ...
 $ Ratio6: num  151.14 5.21 383.2 67.61 110.99 ...
 $ Ratio7: num  67.7 28.1 794 238.1 230.5 ...

str(firms.non.failed)

'data.frame':   30 obs. of  7 variables:
 $ Ratio1: num  0.68 0.548 0.057 0.48 0.227 ...
 $ Ratio2: num  5.98 2.68 0.55 1.43 1.27 ...
 $ Ratio3: num  6.25 2.32 2.13 1.21 3.29 ...
 $ Ratio4: num  0.04801 0.02515 0.00211 -0.11028 0.05752 ...
 $ Ratio5: num  0.0354 0.0178 0.0111 -0.0547 0.0633 ...
 $ Ratio6: num  73.2 588.8 54.7 277.4 43.5 ...
 $ Ratio7: num  13.3 177.4 202.8 757 50.3 ...
```

```
summary(firms)
```

```
     Ratio1            Ratio2             Ratio3             Ratio4
Min.   :0.01051   Min.   : 0.01901   Min.   :-0.3294   Min.   :-2.528084
1st Qu.:0.08385   1st Qu.: 0.27187   1st Qu.: 0.1079   1st Qu.:-0.062512
Median :0.15029   Median : 0.75076   Median : 0.4476   Median : 0.004135
Mean   :0.23450   Mean   : 1.58234   Mean   : 2.8446   Mean   : 0.090600
3rd Qu.:0.37674   3rd Qu.: 1.30391   3rd Qu.: 2.1738   3rd Qu.: 0.076268
Max.   :0.68018   Max.   :10.72871   Max.   :37.1129   Max.   : 6.197973

     Ratio5              Ratio6             Ratio7              Class
Min.   :-0.2662416   Min.   :-317.29   Min.   :   7.14   Failed    :10
1st Qu.:-0.0200257   1st Qu.:  42.95   1st Qu.:  33.97   Non-Failed:30
Median : 0.0182522   Median :  77.63   Median :  83.73
Mean   : 0.0007162   Mean   : 196.04   Mean   : 197.91
3rd Qu.: 0.0392552   3rd Qu.: 156.23   3rd Qu.: 232.37
Max.   : 0.0886270   Max.   :3346.04   Max.   :1186.19
```

```
summary(firms.failed)
```

```
     Ratio1            Ratio2            Ratio3             Ratio4
Min.   :0.01051   Min.   :0.01901   Min.   :-0.32942   Min.   :-1.69010
1st Qu.:0.06116   1st Qu.:0.15168   1st Qu.:-0.05306   1st Qu.:-0.09581
Median :0.12376   Median :0.22810   Median : 0.15459   Median :-0.03524
Mean   :0.12724   Mean   :0.36280   Mean   : 0.21331   Mean   : 0.46977
3rd Qu.:0.15375   3rd Qu.:0.43300   3rd Qu.: 0.26694   3rd Qu.: 0.01874
Max.   :0.36425   Max.   :1.10630   Max.   : 1.38270   Max.   : 6.19797

     Ratio5             Ratio6            Ratio7
Min.   :-0.26624   Min.   :  3.176   Min.   : 27.18
1st Qu.:-0.06995   1st Qu.: 69.768   1st Qu.: 63.48
Median :-0.03613   Median :125.428   Median :234.29
Mean   :-0.06281   Mean   :139.507   Mean   :240.13
3rd Qu.:-0.01608   3rd Qu.:148.961   3rd Qu.:315.15
Max.   : 0.02031   Max.   :383.200   Max.   :793.95
```

```
summary(firms.non.failed)
```

```
     Ratio1            Ratio2             Ratio3             Ratio4
Min.   :0.02010   Min.   : 0.1153   Min.   :-0.1752   Min.   :-2.52808
1st Qu.:0.08972   1st Qu.: 0.5421   1st Qu.: 0.2127   1st Qu.:-0.02077
Median :0.25772   Median : 0.9459   Median : 1.2444   Median : 0.02601
Mean   :0.27026   Mean   : 1.9889   Mean   : 3.7217   Mean   :-0.03579
3rd Qu.:0.41478   3rd Qu.: 1.5296   3rd Qu.: 2.8467   3rd Qu.: 0.07636
Max.   :0.68018   Max.   :10.7287   Max.   :37.1129   Max.   : 1.07696

     Ratio5              Ratio6             Ratio7
Min.   :-0.13056   Min.   :-317.29   Min.   :   7.14
1st Qu.: 0.01021   1st Qu.:  41.85   1st Qu.:  28.33
Median : 0.02818   Median :  69.79   Median :  77.72
Mean   : 0.02189   Mean   : 214.89   Mean   : 183.84
3rd Qu.: 0.04603   3rd Qu.: 158.54   3rd Qu.: 190.64
Max.   : 0.08863   Max.   :3346.04   Max.   :1186.19
```

```
# Observation matrix (X), Centered matrix (X0), Covariance matrix (S) and Correlation
# matrix (R)

firms.obs<-firms[,-c(8)]
X<-as.matrix(firms.obs)
X0<-scale(X, scale=FALSE)
S<-cov(firms.obs)
R<-cor(firms.obs)


# In order to see if there exists any correlation between ratios, we first compute
# correlation matrices for the two groups

R.f<-cor(firms.failed)
R.nf<-cor(firms.non.failed)
```

```
round(R.f,2)
```

	Ratio1	Ratio2	Ratio3	Ratio4	Ratio5	Ratio6	Ratio7
Ratio1	1.00	0.88	−0.02	−0.21	−0.04	−0.35	−0.15
Ratio2	0.88	1.00	0.01	−0.25	−0.05	−0.37	−0.19
Ratio3	−0.02	0.01	1.00	−0.02	0.56	−0.19	0.22
Ratio4	−0.21	−0.25	−0.02	1.00	0.34	−0.22	0.04
Ratio5	−0.04	−0.05	0.56	0.34	1.00	−0.36	0.49
Ratio6	−0.35	−0.37	−0.19	−0.22	−0.36	1.00	0.55
Ratio7	−0.15	−0.19	0.22	0.04	0.49	0.55	1.00

```
round(R.nf,2)
```

	Ratio1	Ratio2	Ratio3	Ratio4	Ratio5	Ratio6	Ratio7
Ratio1	1.00	0.46	0.07	0.01	0.03	0.23	0.02
Ratio2	0.46	1.00	0.81	0.16	0.26	0.00	0.16
Ratio3	0.07	0.81	1.00	0.14	0.18	−0.05	0.21
Ratio4	0.01	0.16	0.14	1.00	0.16	−0.01	−0.20
Ratio5	0.03	0.26	0.18	0.16	1.00	−0.06	−0.26
Ratio6	0.23	0.00	−0.05	−0.01	−0.06	1.00	0.15
Ratio7	0.02	0.16	0.21	−0.20	−0.26	0.15	1.00

```
# An auxiliary function to compute the rank of a matrix

mrank<-function(a,eps=1.0e-5){
        s<-svd(a)
        r<-sum(s$d>eps)
        return(r)
        }

# Compute ranks of both correlation matrices (R.f and R.nf)

mrank(R.f,eps=1.e-3)
 [1] 7
mrank(R.f,eps=1.e-4)
 [1] 7
mrank(R.nf,eps=1.e-4)
 [1] 7
mrank(R.nf,eps=1.e-3)
 [1] 7

# Results show that both matrices have maximum rank, meaning that in theory there are no
# pairs of ratios perfectly correlated.
# Alternatively, we compute condition numbers to obtain a similar result.

kappa(R.f)
 [1] 78.96666
kappa(R.nf)
 [1] 24.71842

# A description of correlations by group

V.f<-R.f[lower.tri(R.f)]
V.nf<-R.nf[lower.tri(R.nf)]

hist(V.f,nclass=8)
hist(V.nf,nclass=8)

# Apply LDA without function 'lda' directly
# So, calculate difference of sample means and sample covariance matrix, etc

m.f<-apply(firms.failed,2,mean)
m.nf<-apply(firms.non.failed,2,mean)
d<-m.f-m.nf

S.f<-cov(firms.failed)
S.nf<-cov(firms.non.failed)

n1<-nrow(firms.failed)
n2<-nrow(firms.non.failed)
n<-n1+n2

W.f<-(n1-1)*S.f
```

```
W.nf<-(n2-1)*S.nf
W<-W.f+W.nf
S<-(1/(n-2))*W

S1<-solve(S)

L<-S1%*%d
L.norm<-L/L[7]
L.norm

                [,1]
Ratio1  -4.345902e+03
Ratio2   2.421489e+02
Ratio3  -1.072329e+02
Ratio4   5.915756e+02
Ratio5  -2.459089e+04
Ratio6  -1.983339e-01
Ratio7   1.000000e+00

# Apply LDA with function 'lda' directly, using Leave-one-out (LOO) technique

require(MASS)

firms.lda.1<-lda(Class~.,data=firms,CV=TRUE)

# Using LOO, R does not provide the coefficients. If we explicitly want them,
# we must apply LDA without using LOO

firms.lda.2<-lda(Class~.,data=firms,CV=FALSE)

Coefficients of linear discriminants:
               LD1
Ratio1   2.762364153
Ratio2  -0.153915869
Ratio3   0.068159897
Ratio4  -0.376020250
Ratio5  15.630587085
Ratio6   0.000126066
Ratio7  -0.000635625

# Normalising the vector by Ratio7, we obtain the same coefficients

# We compute Confusion matrix and determine the errors to assess LDA performance using LOO

Observed<-firms$Class
n<-length(Observed)
n
[1] 40

Predicted<-rep("Non-Failed",n)
Predicted[firms.lda.1$posterior[1,]>0.5]<-"Failed"

Confusion.matrix.lda.1<-table(Observed,Predicted)
Confusion.matrix.lda.1

             Predicted
 Observed     Failed Non-Failed
   Failed         10          0
   Non-Failed     10         20

Err1<-Confusion.matrix.lda.1[1,2]/sum(Confusion.matrix.lda.1[1,])
Err1
 [1] 0
Err2<-Confusion.matrix.lda.1[2,1]/sum(Confusion.matrix.lda.1[2,])
Err2
 [1] 0.3333333
Overall.Err<-1-sum(diag(Confusion.matrix.lda.1))/sum(Confusion.matrix.lda.1)
Overall.Err
 [1] 0.25
```

```
#
# Linear Discriminant Analysis (LDA)
#
# Sample: 66 American manufacturing firms (33 Bankrupt and 33 Solvent)
#
# Data from the 1945-1965 period (yearly)
#

# Read data table, stored in "Altman.data.txt" for example

altman.firms<-read.table("Altman.data.txt",header=TRUE,row.names=1)

altman.firms.b<-altman.firms[altman.firms$Class=="B",-c(1)]
altman.firms.s<-altman.firms[altman.firms$Class=="S",-c(1)]

# Table details and descriptive statistics

str(altman.firms)

'data.frame':   66 obs. of  6 variables:
 $ Class: Factor w/ 2 levels "B","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ X1   : num  36.7 24 -61.6 -1 18.9 -57.2 3 -5.1 17.9 5.4 ...
 $ X2   : num  -62.8 3.3 -120.8 -18.1 -3.8 ...
 $ X3   : num  -89.5 -3.5 -103.2 -28.8 -50.6 ...
 $ X4   : num  54.1 20.9 24.7 36.2 26.4 11 8 6.5 22.6 23.8 ...
 $ X5   : num  1.7 1.1 2.5 1.1 0.9 1.7 1 0.5 1 1.5 ...

names(altman.firms)

[1] "Class" "X1"    "X2"    "X3"    "X4"    "X5"

str(altman.firms.b)

'data.frame':   33 obs. of  5 variables:
 $ X1: num  36.7 24 -61.6 -1 18.9 -57.2 3 -5.1 17.9 5.4 ...
 $ X2: num  -62.8 3.3 -120.8 -18.1 -3.8 ...
 $ X3: num  -89.5 -3.5 -103.2 -28.8 -50.6 ...
 $ X4: num  54.1 20.9 24.7 36.2 26.4 11 8 6.5 22.6 23.8 ...
 $ X5: num  1.7 1.1 2.5 1.1 0.9 1.7 1 0.5 1 1.5 ...

str(altman.firms.s)

'data.frame':   33 obs. of  5 variables:
 $ X1: num  35.2 38.8 14 55.1 59.3 33.6 52.8 45.6 47.4 40 ...
 $ X2: num  43 47 -3.3 35 46.7 20.8 33 26.1 68.6 37.3 ...
 $ X3: num  16.4 16 4 20.8 12.6 12.5 23.6 10.4 13.8 33.4 ...
 $ X4: num  99.1 126.5 91.7 72.3 724.1 ...
 $ X5: num  1.3 1.9 2.7 1.9 0.9 2.4 1.5 2.1 1.6 3.5 ...

summary(altman.firms)

 Class         X1                  X2                  X3
 B:33   Min.   :-185.100   Min.   :-308.90   Min.   :-280.000
 S:33   1st Qu.:   5.175   1st Qu.: -39.05   1st Qu.: -17.675
        Median :  24.550   Median :   7.85   Median :   4.100
        Mean   :  17.669   Mean   : -13.63   Mean   :  -8.226
        3rd Qu.:  45.975   3rd Qu.:  35.75   3rd Qu.:  14.400
        Max.   :  72.400   Max.   :  68.60   Max.   :  34.100
       X4                 X5
 Min.   :  0.70   Min.   :0.100
 1st Qu.: 21.93   1st Qu.:1.025
 Median : 86.00   Median :1.550
 Mean   :147.35   Mean   :1.721
 3rd Qu.:214.00   3rd Qu.:1.975
 Max.   :771.70   Max.   :6.700

summary(altman.firms.b)
```

```
       X1                  X2                  X3                  X4
 Min.   :-185.100   Min.   :-308.90   Min.   :-280.00   Min.   :  0.70
 1st Qu.: -17.200   1st Qu.: -98.00   1st Qu.: -35.70   1st Qu.: 11.00
 Median :   5.100   Median : -39.40   Median : -17.70   Median : 21.70
 Mean   :  -6.047   Mean   : -62.51   Mean   : -31.77   Mean   : 40.05
 3rd Qu.:  18.900   3rd Qu.: -13.10   3rd Qu.:  -6.50   3rd Qu.: 35.70
 Max.   :  72.400   Max.   :  20.80   Max.   :   6.80   Max.   :267.90
       X5
 Min.   :0.100
 1st Qu.:0.900
 Median :1.200
 Mean   :1.503
 3rd Qu.:1.700
 Max.   :6.700

summary(altman.firms.s)

       X1                  X2                  X3                  X4
 Min.   :14.00   Min.   :-3.30   Min.   :-14.40   Min.   : 53.4
 1st Qu.:33.80   1st Qu.:21.50   1st Qu.:  7.10   1st Qu.:105.6
 Median :45.60   Median :35.90   Median : 14.60   Median :164.4
 Mean   :41.38   Mean   :35.25   Mean   : 15.32   Mean   :254.7
 3rd Qu.:52.80   3rd Qu.:47.00   3rd Qu.: 23.60   3rd Qu.:307.5
 Max.   :69.00   Max.   :68.60   Max.   : 34.10   Max.   :771.7
       X5
 Min.   :0.900
 1st Qu.:1.500
 Median :1.800
 Mean   :1.939
 3rd Qu.:2.000
 Max.   :5.500

# Observation matrix (X), Centered matrix (X0), Covariance matrix (S) and Correlation
# matrix (R)

altman.firms.obs<-altman.firms[,-c(1)]
altman.X<-as.matrix(altman.firms.obs)
altman.X0<-scale(atman.X, scale=FALSE)
altman.S<-cov(altman.firms.obs)
altman.R<-cor(altman.firms.obs)


# In order to see if there exists any correlation between ratios, we first compute
# correlation matrices for the two groups

altman.R.b<-cor(altman.firms.b)
altman.R.s<-cor(altman.firms.s)

round(altman.R.b,2)

      X1    X2    X3   X4    X5
X1  1.00  0.60 -0.13 0.12  0.33
X2  0.60  1.00  0.45 0.05 -0.19
X3 -0.13  0.45  1.00 0.06 -0.78
X4  0.12  0.05  0.06 1.00  0.03
X5  0.33 -0.19 -0.78 0.03  1.00

round(altman.R.s,2)

     X1   X2   X3    X4    X5
X1 1.00 0.50 0.11  0.33  0.15
X2 0.50 1.00 0.29  0.48  0.06
X3 0.11 0.29 1.00  0.36  0.27
X4 0.33 0.48 0.36  1.00 -0.08
X5 0.15 0.06 0.27 -0.08  1.00

# An auxiliary function to compute the rank of a matrix

mrank<-function(a,eps=1.0e-5){
        s<-svd(a)
        r<-sum(s$d>eps)
        return(r)
        }
```

```
# Compute ranks of both correlation matrices (altman.R.b and altman.R.s)

mrank(altman.R.b,eps=1.e-3)
 [1] 5
mrank(altman.R.b,eps=1.e-4)
 [1] 5
mrank(altman.R.s,eps=1.e-4)
 [1] 5
mrank(altman.R.s,eps=1.e-3)
 [1] 5

# Results show that both matrices have maximum rank, meaning that in theory there are no
# pairs of ratios perfectly correlated.
# Alternatively, we compute condition numbers to obtain a similar result.

kappa(altman.R.b)
 [1] 16.25495
kappa(altman.R.s)
 [1] 6.016638

# A description of correlations by group

altman.V.b<-altman.R.b[lower.tri(altman.R.b)]
altman.V.s<-altman.R.s[lower.tri(altman.R.s)]

hist(altman.V.b,nclass=8)
hist(altman.V.nf,nclass=8)

# Apply LDA without function 'lda' directly
# So, calculate difference of sample means and sample covariance matrix, etc

m.b<-apply(altman.firms.b,2,mean)
m.s<-apply(altman.firms.s,2,mean)
d<-m.b-m.s

S.b<-cov(altman.firms.b)
S.s<-cov(altman.firms.s)

n1<-nrow(altman.firms.b)
n2<-nrow(altman.firms.s)
n<-n1+n2

S<-(S.b+S.s)/2

S1<-solve(S)

L<-S1%*%d
L.norm<-L/L[5]
round(L.norm,3)

     [,1]
X1 0.012
X2 0.015
X3 0.033
X4 0.006
X5 1.000

# Apply LDA with function 'lda' directly, using Leave-one-out (LOO) technique

require(MASS)

altman.firms.lda.1<-lda(Class~.,data=altman.firms,CV=TRUE)

# Using LOO, R does not provide the coefficients. If we explicitly want them,
# we must apply LDA without using LOO

altman.firms.lda.2<-lda(Class~.,data=altman.firms,CV=FALSE)
```

```
Coefficients of linear discriminants:
           LD1
X1 0.005948816
X2 0.007063952
X3 0.016180383
X4 0.002977887
X5 0.485106138

# Normalising the vector by X5, we obtain the same coefficients

# We compute Confusion matrix and determine the errors to assess LDA performance using LOO

altman.Observed<-altman.firms$Class
altman.n<-length(altman.Observed)
altman.n
[1] 66
altman.Predicted<-rep("S",altman.n)
altman.Predicted[altman.firms.lda.1$posterior[1,]>0.5]<-"B"

altman.Confusion.matrix.lda.1<-table(altman.Observed,altman.Predicted)
altman.Confusion.matrix.lda.1

                altman.Predicted
altman.Observed  B  S
              B 17 16
              S 16 17

altman.Err1<-altman.Confusion.matrix.lda.1[1,2]/sum(altman.Confusion.matrix.lda.1[1,])
altman.Err1
 [1] 0.4848485
altman.Err2<-altman.Confusion.matrix.lda.1[2,1]/sum(altman.Confusion.matrix.lda.1[2,])
altman.Err2
 [1] 0.4848485
altman.Overall.Err<-1-sum(diag(altman.Confusion.matrix.lda.1))/sum(altman.Confusion.matrix.lda.1)
altman.Overall.Err
 [1] 0.4848485
```

A.2 Logistic Regression Script

```
# Logistic Regression
#
# Sample: 40 firms (10 Failed and 30 Non-Failed) belonging to the sector
# 01. Crop and animal production, hunting and related service activities (according to
# NACE Rev. 2)
#
# Firms are either SA or SL, indistinctly
# Data from the 2007-2014 period (yearly)
#

# Read data table, stored in "TFG_Observations.txt" for example

firms<-read.table("TFG_Observations.txt",header=TRUE,row.names=1)

# Fit a Logistic Regression model. See 'TFG.LDA.r' for any previous details

y<-(firms$Class=="Failed")*1
firmsn<-data.frame(cbind(firms[,-8],y))

firmsn.glm1<-glm(y~.,family=binomial(link=logit),data=firmsn)
summary(firmsn.glm1)

Call:
  glm(formula = y ~ ., family = binomial(link = logit), data = firmsn)

Deviance Residuals:
   Min          1Q      Median          3Q         Max
-1.40810   -0.37743   -0.07248    0.00528    2.13028

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.580475   1.235284   0.470   0.6384
Ratio1       -8.844292   6.606646  -1.339   0.1807
Ratio2       -0.784971   1.891081  -0.415   0.6781
Ratio3       -2.085571   1.254793  -1.662   0.0965 .
Ratio4        0.912546   1.250231   0.730   0.4655
Ratio5      -25.682263  15.232543  -1.686   0.0918 .
Ratio6        0.000367   0.004273   0.086   0.9316
Ratio7        0.005717   0.003862   1.480   0.1388
---
  Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 44.987  on 39  degrees of freedom
Residual deviance: 18.607  on 32  degrees of freedom
AIC: 34.607

Number of Fisher Scoring iterations: 9

# Stepwise predictor selection involving Akaike's Information Criterium (AIC)

step(firmsn.glm1)

Start:  AIC=34.61
y ~ Ratio1 + Ratio2 + Ratio3 + Ratio4 + Ratio5 + Ratio6 + Ratio7

Df Deviance    AIC
- Ratio6  1   18.614 32.614
- Ratio2  1   18.786 32.786
- Ratio4  1   19.416 33.416
<none>        18.607 34.607
- Ratio1  1   20.782 34.782
- Ratio7  1   21.553 35.553
- Ratio3  1   23.115 37.115
- Ratio5  1   23.675 37.675

Step:  AIC=32.61
y ~ Ratio1 + Ratio2 + Ratio3 + Ratio4 + Ratio5 + Ratio7
```

```
Df Deviance    AIC
- Ratio2   1   18.786 30.786
- Ratio4   1   19.416 31.416
<none>         18.614 32.614
- Ratio1   1   20.783 32.783
- Ratio7   1   21.808 33.808
- Ratio3   1   23.327 35.327
- Ratio5   1   23.700 35.700

Step:  AIC=30.79
y ~ Ratio1 + Ratio3 + Ratio4 + Ratio5 + Ratio7

Df Deviance    AIC
- Ratio4   1   19.781 29.781
<none>         18.786 30.786
- Ratio7   1   22.938 32.938
- Ratio1   1   24.320 34.320
- Ratio3   1   24.323 34.323
- Ratio5   1   24.633 34.633

Step:  AIC=29.78
y ~ Ratio1 + Ratio3 + Ratio5 + Ratio7

Df Deviance    AIC
<none>         19.781 29.781
- Ratio7   1   23.747 31.747
- Ratio5   1   24.714 32.714
- Ratio1   1   25.825 33.825
- Ratio3   1   27.373 35.373

Call:  glm(formula = y ~ Ratio1 + Ratio3 + Ratio5 + Ratio7, family = binomial(link = logit),
           data = firmsn)

Coefficients:
  (Intercept)        Ratio1        Ratio3        Ratio5        Ratio7
0.611688    -10.575905     -2.449561    -26.579763      0.006001

Degrees of Freedom: 39 Total (i.e. Null);  35 Residual
Null Deviance:      44.99
Residual Deviance: 19.78        AIC: 29.78
There were 21 warnings (use warnings() to see them)

# Now we can fit the definitive resulting model. Note that Ratio2, Ratio4 and Ratio6
# have been removed

firmsn.glm2<-glm(y ~ Ratio1 + Ratio3 + Ratio5 + Ratio7,family=binomial(link=logit),data=firmsn)
summary(firmsn.glm2)

Call:
  glm(formula = y ~ Ratio1 + Ratio3 + Ratio5 + Ratio7, family = binomial(link = logit),
      data = firmsn)

Deviance Residuals:
   Min          1Q      Median          3Q         Max
-1.21120   -0.40097   -0.06567    0.00861    2.27194

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.611688   1.046888   0.584   0.5590
Ratio1      -10.575905   5.755136  -1.838   0.0661 .
Ratio3       -2.449561   1.265132  -1.936   0.0528 .
Ratio5      -26.579763  16.357801  -1.625   0.1042
Ratio7        0.006001   0.003774   1.590   0.1118
---
  Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 44.987  on 39  degrees of freedom
Residual deviance: 19.781  on 35  degrees of freedom
AIC: 29.781

Number of Fisher Scoring iterations: 9
```

```
# Confusion matrix (with plug-in predictions)

firmsn.noclass<-firmsn[,-8]
firmsn.glm2.pred<-predict(firmsn.glm2,newdata=firmsn.noclass,type="response")

Observed<-firms$Class
n<-length(Observed)
Predicted<-rep("Non-Failed",n)
Predicted[firmsn.glm2.pred>0.5]<-"Failed"

Confusion.matrix.glm2<-table(Observed,Predicted)
Confusion.matrix.glm2

             Predicted
 Observed     Failed Non-Failed
 Failed            7          3
 Non-Failed        1         29

# An auxiliary function to compute misclassification errors from a [2,2] confusion matrix

misclassification.errors<-function(T){
        Err1<-T[1,2]/sum(T[1,])
        Err2<-T[2,1]/sum(T[2,])
        Overall.Err<-1-sum(diag(T))/sum(T)
        return(list(Err1=Err1,Err2=Err2,Overall.Err=Overall.Err))
        }

# And another small one for syntactic purposes

round2<-function(x){round(x,2)}

# We determine the errors to assess Logistic Regression performance on this specific
# dataset

L<-misclassification.errors(Confusion.matrix.glm2)
lapply(L,round2)

 $Err1
 [1] 0.3

 $Err2
 [1] 0.03

 $Overall.Err
 [1] 0.1

# Finally, we compute Confusion matrix and errors (misclassification rate) again using
# Leave-One-Out method

Observed<-firms$Class
n<-length(Observed)

Predicted<-rep("Non-Failed",n)

for (i in 1:n){
        Train.i<-firmsn[-i,]
        Test.i<-firmsn[i,-8]
        Model.i<-glm(y ~ Ratio1 + Ratio3 + Ratio5 + Ratio7,family=binomial(link=logit),data=Train.i)
        Pred.i<-predict(Model.i,newdata=Test.i,type="response")
        if (Pred.i>0.5) Predicted[i]<- "Failed"
}

Confusion.matrix.glm2.LOO<-table(Observed,Predicted)
Confusion.matrix.glm2.LOO
             Predicted
 Observed     Failed Non-Failed
  Failed           6          4
  Non-Failed       3         27

L<-misclassification.errors(Confusion.matrix.glm2.LOO)
lapply(L,round2)
```

```
$Err1
[1] 0.4

$Err2
[1] 0.1

$Overall.Err
[1] 0.18
```

A.3 KNN Script

```
#
# k-Nearest Neighbours (KNN)
#
# Sample: 40 firms (10 Failed and 30 Non-Failed) belonging to the sector
# 01. Crop and animal production, hunting and related service activities (according to
# NACE Rev. 2)
#
# Firms are either SA or SL, indistinctly
# Data from the 2007-2014 period (yearly)
#


# Read data table, stored in "TFG_Observations.txt" for example. See 'TFG.LDA.r' for
# any previous details

firms<-read.table("TFG_Observations.txt",header=TRUE,row.names=1)

firms.failed<-firms[firms$Class=="Failed",-8]
firms.non.failed<-firms[firms$Class=="Non-Failed",-8]

# An auxiliary function to compute misclassification errors from a [2,2] confusion matrix

misclassification.errors<-function(T){
        Err1<-T[1,2]/sum(T[1,])
        Err2<-T[2,1]/sum(T[2,])
        Overall.Err<-1-sum(diag(T))/sum(T)
        return(list(Err1=Err1,Err2=Err2,Overall.Err=Overall.Err))
        }

# And another small one for syntactic purposes

round2<-function(x){round(x,2)}

# Classify using k-Nearest Neighbours and LOO with k = 1

require(class)

X<-firms[,1:7]
y<-firms$Class

Observed<-firms$Class
n<-length(Observed)
n
 [1] 40

k<-1

firms.knn.1<-knn.cv(X,y,k,prob=TRUE)

Predicted<-firms.knn.1

# Confusion matrix (with plugin predictions) and misclassification rate for k = 1

Confusion.matrix.knn.1<-table(Observed,Predicted)
Confusion.matrix.knn.1

             Predicted
 Observed     Failed Non-Failed
   Failed          5          5
   Non-Failed      4         26

L<-misclassification.errors(Confusion.matrix.knn.1)
lapply(L,round2)

 $Err1
 [1] 0.5

 $Err2
 [1] 0.13
```

```
 $Overall.Err
 [1] 0.22

# A function to classify in terms of k. Results for k = 1,2,3

do.knn<-function(k){
        firms.knn.k<-knn.cv(X,y,k,prob=TRUE)
        Predicted<-firms.knn.k
        Confusion.matrix.knn.k<-table(Observed,Predicted)
        print(Confusion.matrix.knn.k)
        L<-misclassification.errors(Confusion.matrix.knn.k)
        print(lapply(L,round2))
        }

do.knn(1)
            Predicted
Observed     Failed Non-Failed
  Failed          5          5
  Non-Failed      4         26

$ Err1
[1] 0.5

$ Err2
[1] 0.13

$ Overall.Err
[1] 0.22

do.knn(2)

            Predicted
Observed     Failed Non-Failed
  Failed          4          6
  Non-Failed      5         25

$ Err1
[1] 0.6

$ Err2
[1] 0.17

$ Overall.Err
[1] 0.28

do.knn(3)

            Predicted
Observed     Failed Non-Failed
  Failed          4          6
  Non-Failed      2         28

$ Err1
[1] 0.6

$ Err2
[1] 0.07

$ Overall.Err
[1] 0.2
```

A.4 CART Script

```
#
# Classification and Regression Trees (CART)
#
# Sample: 40 firms (10 Failed and 30 Non-Failed) belonging to the sector
# 01. Crop and animal production, hunting and related service activities (according to
# NACE Rev. 2)
#
# Firms are either SA or SL, indistinctly
# Data from the 2007-2014 period (yearly)
#

# Read data table, stored in "TFG_Observations.txt" for example. See 'TFG.LDA.r' for
# any previous details

firms<-read.table("TFG_Observations.txt",header=TRUE,row.names=1)

firms.failed<-firms[firms$Class=="Failed",-8]
firms.non.failed<-firms[firms$Class=="Non-Failed",-8]

# An auxiliary function to compute misclassification errors from a [2,2] confusion matrix

misclassification.errors<-function(T){
        Err1<-T[1,2]/sum(T[1,])
        Err2<-T[2,1]/sum(T[2,])
        Overall.Err<-1-sum(diag(T))/sum(T)
        return(list(Err1=Err1,Err2=Err2,Overall.Err=Overall.Err))
        }

# And another small one for syntactic purposes

round2<-function(x){round(x,2)}

# Build a Classification Tree using Gini Impurity Measure

require(rpart)
require(rpart.plot)

firms.tree<-rpart(Class~., data=firms, method="class")
rpart.plot(firms.tree)
summary(firms.tree)

# Build a Classification Tree using Entropy

firms.tree.2<-rpart(Class~., data=firms, method="class", minsplit=10,
                    parms=list(split="information"))
rpart.plot(firms.tree.2)
summary(firms.tree.2)

#
# Confusion matrix and misclassification rate in both cases
#

Predicted<-predict(firms.tree, firms, type="class")
Confusion.matrix.tree<-table(Observed=firms$Class, Predicted = Predicted)
as.matrix(Confusion.matrix.tree)

             Predicted
 Observed     Failed Non-Failed
   Failed          8          2
   Non-Failed      5         25

L<-misclassification.errors(Confusion.matrix.tree)
lapply(L,round2)

$Err1
[1] 0.2

$Err2
[1] 0.17
```

```
$Overall.Err
[1] 0.18

Predicted.2<-predict(firms.tree.2, firms, type="class")
Confusion.matrix.tree.2<-table(Observed=firms$Class, Predicted = Predicted.2)
as.matrix(Confusion.matrix.tree.2)

             Predicted
 Observed     Failed Non-Failed
   Failed          8          2
   Non-Failed      1         29

L.2<-misclassification.errors(Confusion.matrix.tree.2)
lapply(L.2,round2)

$Err1
[1] 0.2

$Err2
[1] 0.03

$Overall.Err
[1] 0.07
```

A.5 Neural Networks Script

```
#
#
# Neural Networks (NN)
#
# Sample: 40 firms (10 Failed and 30 Non-Failed) belonging to the sector
# 01. Crop and animal production, hunting and related service activities (according to
# NACE Rev. 2)
#
# Firms are either SA or SL, indistinctly
# Data from the 2007-2014 period (yearly)
#

# Read data table, stored in "TFG_Observations.txt" for example. See 'TFG.LDA.r' for
# any previous details

firms<-read.table("TFG_Observations.txt",header=TRUE,row.names=1)

firms.failed<-firms[firms$Class=="Failed",-8]
firms.non.failed<-firms[firms$Class=="Non-Failed",-8]

# An auxiliary function to compute misclassification errors from a [2,2] confusion matrix

misclassification.errors<-function(T){
        Err1<-T[1,2]/sum(T[1,])
        Err2<-T[2,1]/sum(T[2,])
        Overall.Err<-1-sum(diag(T))/sum(T)
        return(list(Err1=Err1,Err2=Err2,Overall.Err=Overall.Err))
        }

# And another small one for syntactic purposes

round2<-function(x){round(x,2)}

# Classification using the multilayer perceptron with 2 layers
# and 2 units in the hidden layer

require(nnet)

firms.nn<-nnet(Class~., data=firms, size=2)
summary(firms.nn)
print(firms.nn)

# Confusion matrix and misclassification rate

Observed<-firms$Class
n<-length(Observed)

Predicted<-predict(firms.nn, firms, type="class")
Confusion.matrix.nn<-table(Observed=firms$Class, Predicted = Predicted)
as.matrix(Confusion.matrix.nn)

           Predicted
Observed    Failed Non-Failed
 Failed         10          0
 Non-Failed      3         27

L<-misclassification.errors(Confusion.matrix.nn)
lapply(L,round2)

$Err1
[1] 0

$Err2
[1] 0.1

$Overall.Err
[1] 0.07

# Classification using the multilayer perceptron with 2 layers
# and 3 units in the hidden layer
```

```
firms.nn.2<-nnet(Class~., data=firms, size=3)
summary(firms.nn.2)
print(firms.nn.2)

# Confusion matrix and misclassification rate

Predicted.2<-predict(firms.nn.2, firms, type="class")
Confusion.matrix.nn.2<-table(Observed=firms$Class, Predicted = Predicted.2)
as.matrix(Confusion.matrix.nn.2)

           Predicted
Observed    Failed Non-Failed
 Failed          8          2
 Non-Failed      7         23

L<-misclassification.errors(Confusion.matrix.nn.2)
lapply(L,round2)

$Err1
[1] 0.2

$Err2
[1] 0.23

$Overall.Err
[1] 0.22
```