UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL


# Modelos de Previsão do Incumprimento de Crédito pelas Médias Empresas

Mariana Rita da Silva Glória Franco

**Mestrado em Matemática Aplicada à Economia e Gestão**

Trabalho de Projeto orientado por:
Professora Doutora Maria Teresa Alpuim

2023

# Agradecimentos

Gostaria de começar por agradecer à professora Teresa Alpuim, orientadora deste projeto, não só por toda a ajuda e interesse demonstrado ao longo deste trabalho, mas também pelo apoio ao longo de toda a minha licenciatura e mestrado.

Ao Sérgio Naito, agradeço pela disponibilidade, pela ajuda a encontrar este tema e por todo o conhecimento transmitido ao longo deste último ano.

Aos meus pais, agradeço por sempre me apoiarem e nunca desistirem de mim, por serem uma força constante na minha vida, estarem lá nos meus melhores momentos e me ampararem nos meus piores. Não estaria onde estou hoje sem vocês.

À minha irmã, agradeço por ser uma constante fonte de alegria para mim e por me ajudar tanto, mesmo não o sabendo, apenas por partilhar a vida comigo.

Às minhas amigas Jodi e Nicole, que, embora não tendo feito parte do meu percurso académico, foram uma força constante durante este tempo e me ouviram sempre que se tornava mais complicado. Obrigada.

Por fim, e com especial destaque, quero agradecer às minhas amigas e companheiras deste percurso, Marta e Sofia. Sem dúvida alguma que estes últimos anos não teriam sido o mesmo sem a vossa amizade, apoio e companheirismo. Começámos juntas e acabamos juntas, apenas para continuar a apoiar-nos noutras fases das nossas vidas. Obrigada por tudo.

# Resumo

Este projeto tem como objetivo desenvolver um ou mais modelos de previsão do incumprimento de uma Média Empresa, de modo a auxiliar o respetivo banco no processo de tomada de decisão de concessão de crédito. Pretende obter-se, para cada cliente, uma probabilidade de incumprimento que permita classificá-lo como cumpridor ou incumpridor.

Sendo o incumprimento um evento binário, pretende-se desenvolver o modelo de previsão com base numa Regressão Logística. Para tal, possui-se um conjunto vasto de informação sobre a situação de empresas em determinados momentos temporais, bem como informação relativa a eventual incumprimento nos 12 meses seguintes. A informação da situação da empresa constitui o conjunto de possíveis variáveis explicativas do modelo a ser criado, sendo esta relativa, quer a informação de balanço da empresa, quer a informação qualitativa. A variável resposta corresponderá à variável binária de incumprimento já referida, pretendendo-se prever o incumprimento a 12 meses.

Depois de um tratamento prévio às diferentes variáveis explicativas, desde a eliminação de variáveis bastante correlacionadas com outras, até ao agrupamento dos diferentes níveis de algumas variáveis categóricas, obtendo-se assim um número inferior de categorias, com base em diferentes métodos estatísticos, foram construídos diferentes modelos de regressão logística, utilizando diferentes abordagens estatísticas. Embora se tenham obtido modelos com diferentes características, nenhum apresentou resultados significativamente diferentes dos restantes, pelo que qualquer um deles seria utilizável. No entanto, optou-se pela seleção do modelo que, mostrando um bom ajustamento aos dados e uma boa capacidade preditiva, utiliza um menor número de variáveis explicativas.

Em conclusão, o objetivo do projeto foi concretizado com resultados bastante satisfatórios, tendo sido obtidos modelos com elevada qualidade, e gerando-se também previsões muito próximas da realidade ao aplicar o modelo final a uma base de dados diferente da que foi utilizada para o seu desenvolvimento, o que demonstra a adequabilidade do modelo.

**Palavras-Chave:** Probabilidade de Incumprimento; Regressão Logística; Médias Empresas; Análise de Componentes Principais

# Abstract

This project's goal is to develop one or more models to predict the default of a Medium Enterprise, to assist the respective bank in the decision-making process of granting credit. It is intended to obtain, for each client, a probability of default that will allow to classify that client as compliant or non-compliant.

Since the default is a binary event, it is intended to develop the prediction model based on a Logistic Regression. To this end, there is a wide range of information available about the situation of enterprises at some points in time, as well as information about eventual defaults in the following 12 months. The first information mentioned constitutes the set of possible explanatory variables of the model to be created, which are related to both the company's balance sheet information and qualitative information. The response variable will correspond to the binary variable of default already mentioned, with the intention of predicting defaults in the next 12 months.

After a previous treatment of the different explanatory variables, from eliminating highly correlated variables, to grouping different levels of some categorical variables, thus obtaining a lower number of categories, based on different statistical methods, different logistic regression models were constructed, using different statistical approaches. Although there were obtained models with different characteristics, none presented significantly different results from the others, so any of them would be usable. However, the model selected was the one that, showing a good adjustment to the data and predictive capacity, uses a smaller number of explanatory variables.

In conclusion, the project's goal was achieved with very satisfactory results, having been obtained models with high quality, and generating predictions very close to reality by applying the final model to a different database from the one used for its development, which demonstrates the suitability of the model.

**Key Words:** Probability of Default; Logistic Regression; Medium Enterprises; Principal Component Analysis

# Índice

# Índice de Figuras

# Índice de Tabelas

# Lista de Abreviaturas

| | |
|---|---|
| AIC | *Akaike Information Criterion* (Critério de Informação de Akaike) |
| AUC | *Area Under the Curve* (Área abaixo da curva) |
| CP | Componente Principal |
| FN | Falso Negativo |
| FP | Falso Positivo |
| ME | Média(s) Empresa(s) |
| PD | *Probability of Default* (Probabilidade de Incumprimento) |
| PIB | Produto Interno Bruto |
| PME | Pequenas e Médias Empresas |
| ROC | *Receiving Operating Characteristic* |
| ROE | *Return On Equity* |
| VIF | *Variance Inflation Factor* (Fator de Inflação da Variância) |
| VN | Verdadeiro Negativo |
| VP | Verdadeiro Positivo |

# 1. Introdução

O presente estudo foi desenvolvido no âmbito do trabalho de projeto do Curso de Mestrado em Matemática Aplicada à Economia e Gestão.

O principal tema abordado é o risco de incumprimento de crédito, mais precisamente da parte de uma Média Empresa. O presente estudo será baseado no caso de Portugal, com foco específico no caso de um determinado banco português, cujos dados serviram de base para as análises realizadas.

As características analisadas para determinar a probabilidade de incumprimento dependem do tipo de devedor em análise. Enquanto créditos à habitação e créditos individuais a particulares estão intrinsecamente relacionados com as características individuais das pessoas, como a sua idade, situação profissional, entre outras, no caso das empresas é muito importante analisar as suas características do ponto de vista do negócio, pois são estas que podem determinar o possível incumprimento das suas responsabilidades de crédito para com um determinado banco.

É de extrema importância a recolha de informação robusta acerca de uma empresa, neste caso média, para a decisão de concessão de crédito. A informação que servirá de base ao presente estudo destaca-se pela sua completude, transmitindo muito conhecimento acerca de cada empresa apresentada na base de dados. Para além disto, destaca-se ainda o facto de se possuir não só informação relativa ao balanço e condição financeira da empresa, mas também informação qualitativa que permite avaliar, por exemplo, o setor de atividade da mesma, aspetos relativos aos seus sócios e outras características da mesma. A combinação destes dois tipos de informação é determinante, pois permite construir uma descrição o mais completa possível da situação atual da empresa, considerando distintos pontos de vista.

O objetivo deste projeto é construir um modelo de regressão logística múltipla que permita obter, para cada empresa de que se possua a informação necessária, uma probabilidade de incumprimento nos 12 meses seguintes, sendo possível, depois, através dessa probabilidade, classificá-la como futura incumpridora ou não.

Para alcançar o objetivo pretendido, serão utilizadas diversas metodologias estatísticas, quer numa fase inicial, para tratamento de variáveis antes da construção de qualquer modelo, onde são utilizadas metodologias como o Método de Ward e a análise de correlações de Pearson, quer numa segunda fase em que já se construiu um ou mais modelos e se pretende melhorá-los, onde são utlizadas metodologias como o Método de Seleção *Stepwise*, a Análise de Fatores de Inflação da Variância e a Análise de Componentes Principais.

O desafio a cumprir neste projeto será averiguar se os dados disponibilizados para o estudo são suficientes para obter uma previsão de qualidade, que possa ser aplicável a uma Média Empresa que eventualmente pretenda obter um financiamento num banco.

No presente documento é descrito todo o processo desenvolvido ao longo do projeto, iniciando-se com o enquadramento do tema, tanto do ponto de vista dos bancos como credores como do ponto de vista das Médias Empresas como devedoras. De seguida, apresenta-se as metodologias estatísticas utilizadas ao longo do projeto, seguindo-se a apresentação da base de dados utilizada e posterior construção e avaliação dos modelos de regressão. Nos últimos capítulos apresenta-se a interpretação dos modelos do ponto de vista dos diferentes fatores do negócio, a aplicação do modelo selecionado a uma base de dados independente e, por fim, um exemplo de simulação da perda do banco com a utilização do modelo selecionado.

# 2. Enquadramento do Tema

## 2.1.Credores: Bancos

### 2.1.1. Risco de Crédito

O principal negócio de um banco é o de funcionar como um intermediário financeiro, utilizando os recursos dos seus clientes que pretendem guardar, poupar ou investir o seu dinheiro, para fornecer financiamento a clientes que necessitam de recursos. Deste modo, uma parte significativa da atividade bancária funciona através da concessão de empréstimos, existindo vários tipos de créditos, como, por exemplo, individuais, à habitação ou a empresas.

De acordo com a legislação portuguesa, um empréstimo bancário é considerado um mútuo. Um mútuo, conforme partilhado pelo Diário da República Eletrónico, na secção dedicada ao "Lexionário", corresponde a "um contrato que se encontra regulado na lei portuguesa (vide regime geral nos artigos 1142.º e ss. do Código Civil – CC), pelo qual alguém (mutuante) empresta a outrem (mutuário) dinheiro ou outra coisa fungível, ficando o mutuário obrigado a restituir outro tanto do mesmo género ou qualidade.".

Sendo este o seu principal negócio, um dos maiores tipos de risco financeiro a que o setor bancário está exposto é o risco de crédito. Globalmente, o risco de crédito consiste na possibilidade de uma perda por parte da instituição financeira resultante do não reembolso de um empréstimo ou do incumprimento das obrigações contratuais por parte do mutuário. Do ponto de vista do mutuante, este tipo de risco perturba os seus *cash-flows* e aumenta os seus custos de recuperação, que podem incluir diversas despesas, como, por exemplo, a contratação de advogados ou ações judiciais. A perda do banco pode ser parcial ou total, caso o mutuante perca uma parte ou a totalidade do empréstimo concedido ao mutuário, respetivamente.

Contudo, o conceito de risco de crédito pode ser interpretado de uma forma mais abrangente, englobando diferentes medidas de risco conforme exposto por Rajani (2020), nos artigos "The Credit Decision", "Classifications and Key Concepts of Credit Risk (I)" e "Classifications and Key Concepts of Credit Risk (II)". Segundo Rajani, destacam-se os seguintes tipos de risco de crédito:

- Risco de Incumprimento/*Default*: risco de um mutuário não pagar um empréstimo de acordo com os termos do acordo de crédito;
- Risco de Recuperação: risco de uma recuperação inferior à esperada, em caso de incumprimento, causar uma perda mais severa do que a esperada;
- Risco de Exposição: risco de uma exposição superior à esperada, em caso de incumprimento, causar uma perda mais severa do que a esperada;
- Risco de Migração: risco de deterioração da qualidade do crédito e do valor de mercado de um ativo ou posição;
- Risco de *Spread*: risco de alteração dos *spreads* em condições de mercado adversas;
- Risco de Liquidação: risco de deterioração da liquidez e do valor dos ativos durante condições de mercado adversas, reduzindo o seu valor de mercado;
- Risco de Concentração: risco de os mutuários estarem expostos a fatores de risco comuns que podem afetar simultaneamente a sua vontade e capacidade de reembolsar as suas obrigações.

Quanto aos riscos de incumprimento, recuperação e exposição, Rajani resume as fontes e respetivas medidas de risco de crédito na Figura 2.1.

| Source | Measure |
|---|---|
| A borrower will not pay back a loan in accordance with the terms of the credit agreement. | The probability of default (PD), which is the likelihood that a borrower will default. |
| A lower (than expected) recovery at the time of a default causes a more severe loss than expected. | The loss given default (LGD), which represents the likely percentage loss if the borrower does default. |
| A greater than expected exposure at the time of a default causes a severe loss than expected. | Exposure at default (EAD), which can be stated as the nominal amount of the loan or the maximum amount available on a credit line. |

*Figura 2.1 - Fontes e correspondentes medidas de risco de crédito, cf. Rajani (2020) em "The Credit Decision"*

No âmbito do estudo a ser realizado no presente relatório, importa destacar o risco de *default*, que será o alvo da análise apresentada, apresentado na primeira linha da Figura 2.1 e descrito com maior detalhe no capítulo 2.1.1.1.

#### 2.1.1.1. Risco de Incumprimento

Como já referido, o risco de incumprimento é medido através da probabilidade de incumprimento (ou *default*). A probabilidade de *default* (PD), tal como o nome indica, corresponde à probabilidade de um mutuário entrar em incumprimento.

Importa começar por definir o conceito de *default* considerado ao longo deste projeto. De acordo com o Regulamento (UE) N.º 575/2013 do Parlamento Europeu e do Conselho de 26 de junho de 2013 relativo aos requisitos prudenciais das instituições de crédito (*Capital Requirements Regulation*), mais precisamente, na Parte III, Título II, Capítulo 3, Secção 6, Subsecção 2, Artigo 178.º "Incumprimento do devedor":

*"1. Deve considerar-se que se verificou uma situação de incumprimento, no que se refere a um dado devedor, quando se verificar pelo menos uma as seguintes situações:*

*a) A instituição considera que, se não recorrer a medidas como o acionamento das eventuais garantias detidas, existe uma probabilidade reduzida que o devedor cumpra na íntegra as suas obrigações de crédito perante a instituição, a empresa-mãe ou qualquer das suas filiais;*

*b) O devedor regista um atraso superior a 90 dias relativamente a uma obrigação de crédito significativa perante a instituição, a sua empresa-mãe ou qualquer das suas filiais. As autoridades competentes podem substituir os 90 dias por 180 dias para as posições em risco garantidas por bens imóveis destinados à habitação ou por bens imóveis com fins comerciais de PME na categoria de risco sobre a carteira de retalho, bem como para as posições em risco perante entidades do setor público. Os 180 dias não são aplicáveis para efeitos do artigo 36.º, n.º 1, alínea m), ou do artigo 127.º.".*

É esta a definição de *default* utilizada pelo banco cuja base de dados serviu de suporte à realização do presente estudo, tendo sido a variável resposta considerada construída com base nesta definição.

Segundo De Laurentis, Maino e Molteni (2011), a definição de uma medida da qualidade de crédito da contraparte, em particular a probabilidade de incumprimento, é essencial para qualquer tipo abordagem

de gestão do risco de crédito. Nestes termos, a determinação da PD pode ser alcançada através das seguintes alternativas:

- A observação das frequências históricas de incumprimento das classes homogéneas dos mutuários;
- A utilização de ferramentas matemáticas e estatísticas, baseadas em grandes bases de dados. As carteiras de crédito do banco, que possuem milhares de posições observadas no seu comportamento histórico, permitem a aplicação de métodos estatísticos. Os modelos combinam vários tipos de informação num *score* que facilita a atribuição dos mutuários a diferentes classes de risco. Os mesmos modelos permitem uma medição *ex ante* pormenorizada da probabilidade esperada e facilitam a monitorização ao longo do tempo;
- A combinação de abordagens de julgamento e mecânicas (métodos híbridos), em que a classificação automática é gerada por sistemas estatísticos ou numérico e os especialistas corrigem os resultados integrando aspetos qualitativos, com o objetivo de chegar a uma classificação que combine ambas as potencialidades. Mesmo neste caso, a observação histórica, combinada com métodos estatísticos, permite atingir uma probabilidade padrão associada a cada classe de classificação;
- A "extração" da probabilidade implícita de incumprimento embutida nos preços de mercado (títulos e ações), sendo claro que este método só pode ser aplicado a contrapartes cotadas em bolsa nos mercados de capitais próprios ou de valores.

No presente estudo, a abordagem escolhida consiste numa simplificação da segunda alternativa apresentada. Isto é, com base num conjunto considerável de dados e em procedimentos estatísticos, o objetivo é atribuir individualmente uma probabilidade de incumprimento a cada potencial cliente, de modo a que, num momento inicial de possível concessão de crédito, o banco possa decidir atribuir ou não o empréstimo a uma empresa.

## 2.2.Devedores: Empresas

### 2.2.1. Médias Empresas

Segundo informação disponibilizada pela plataforma "Portugal 2020", de acordo com a Recomendação 2003/361/CE, da Comissão, em Portugal uma Média Empresa define-se como uma empresa que emprega entre 51 e 250 pessoas e cujo volume de negócios anual não excede 50 milhões de euros ou o balanço total anual não excede 43 milhões de euros.

De seguida apresentam-se alguns indicadores do contexto de Portugal em relação a este tipo de empresas, obtidas através da plataforma PORDATA. Note-se, porém, que estas análises descritivas englobam dados de empresas que empregam entre 50 a 249 pessoas e cujo volume de negócios anual varia entre 10 e 50 milhões de euros ou cujo balanço total anual varia entre 10 e 43 milhões de euros; a base de dados disponibilizada para o presente estudo considera que uma Média Empresa engloba os clientes empresariais com um volume de faturação entre 1,25 e 50 milhões de euros, pelo que as estatísticas, cuja representação gráfica é apresentada nas figuras 2.2 a 2.9, incluem um menor espectro de empresas, sendo, no entanto, meramente exemplificativas do contexto português.

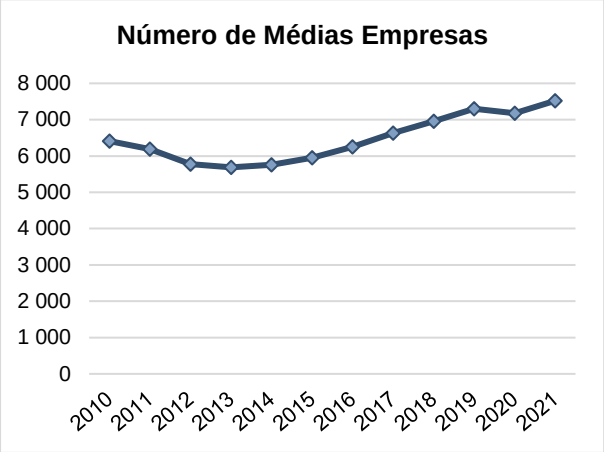


*Figura 2.2 - Número de Médias Empresas em Portugal (2010-2021)*

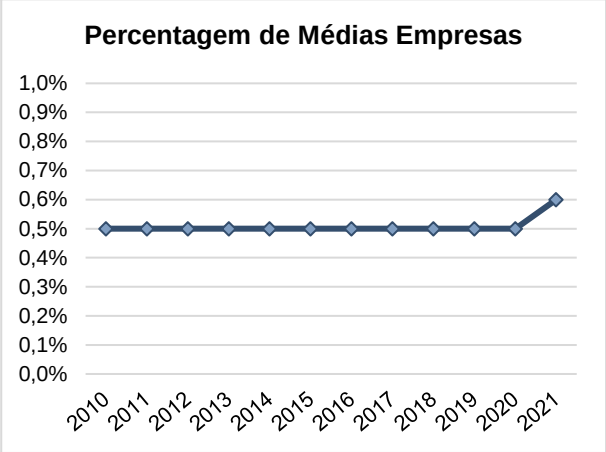


*Figura 2.3 - Percentagem de Médias Empresas em Portugal (2010-2021)*

Enquanto o número de Médias Empresas tem vindo a crescer desde 2013, com um pequeno decréscimo em 2020 e imediata recuperação em 2021, a percentagem deste tipo de empresas no panorama empresarial total português tem-se mantido estável desde 2010 até 2020, com este tipo de empresas a representar apenas 0,5% das empresas no país, aumentado esta percentagem para 0,6% no ano de 2021, o que mostra que o crescimento tem sido global e não apenas ao nível deste tipo de empresas. Como já referido, as empresas incluídas nesta análise descritiva são de um menor espectro que as consideradas no estudo realizado neste projeto, que inclui algumas empresas que seriam consideradas como pequenas nas estatísticas disponibilizadas pela plataforma PORDATA. Se considerarmos o conjunto das pequenas e médias empresas disponibilizadas pela plataforma, a percentagem deste tipo de empresas ascende aos 4%.

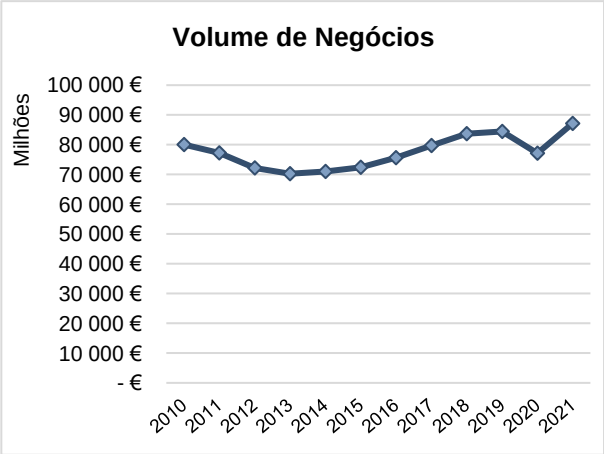


*Figura 2.4 - Volume de Negócios (preços constantes de 2016) das Médias Empresas em Portugal (2010-2021)*

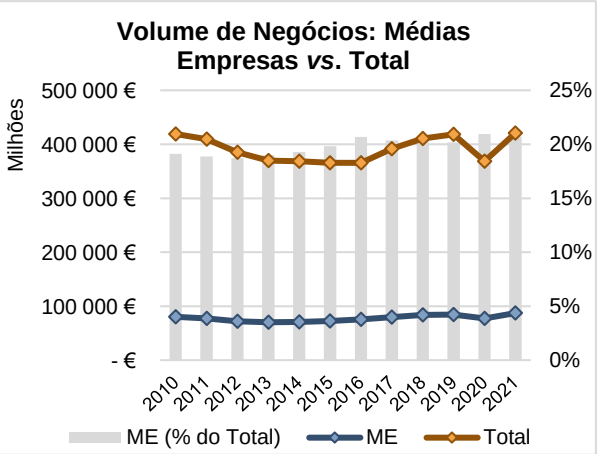


*Figura 2.5- Volume de Negócios (preços constantes de 2016) das Empresas em Portugal: Médias Empresas vs. Total (2010-2021)*

A Figura 2.4 mostra o total do volume de negócios apresentado por estas empresas. Os valores apresentados encontram-se convertidos em preços constantes, possibilitando analisar a evolução dos preços ao longo do tempo, sem o efeito da inflação, determinado a partir da variação dos preços do PIB. Os valores apresentam-se calculados a partir do ano base de 2016, em que o valor a preços constantes coincide com o valor a preços correntes.

Pode verificar-se que, como seria de esperar, tal como o crescimento em termos de número de empresas se verificou desde 2013, o mesmo se verificou com o volume de negócios, observando-se o mesmo decréscimo em 2020 com recuperação em 2021, refletindo o período de início da pandemia de COVID-19. No fim de 2021, o total do volume de negócios das Médias Empresas em Portugal ascendia a quase 88 mil milhões de euros, comparando com cerca de 80 mil milhões em 2010 (verificou-se um crescimento de 8,9%).

Por outro lado, na Figura 2.5 apresenta-se a percentagem do volume de negócios das Médias Empresas relativamente ao total das empresas portuguesas. Observa-se que esta percentagem não varia significativamente ao longo dos anos, sendo que, em média, o volume de negócios das Médias Empresas representa cerca de 19,8% do total do volume de negócios das empresas portuguesas.

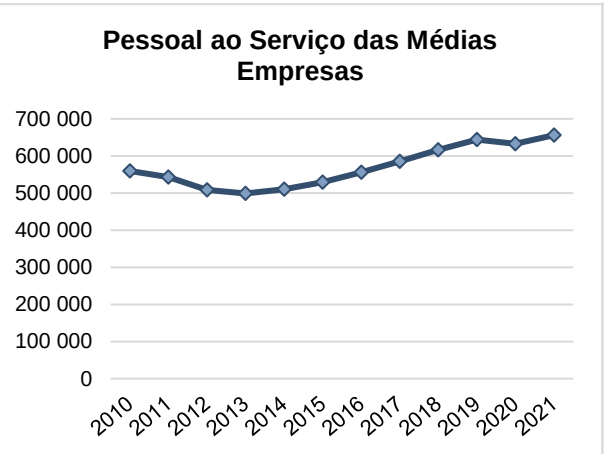


*Figura 2.6 - Pessoal ao Serviço das Médias Empresas em Portugal (2010-2021)*

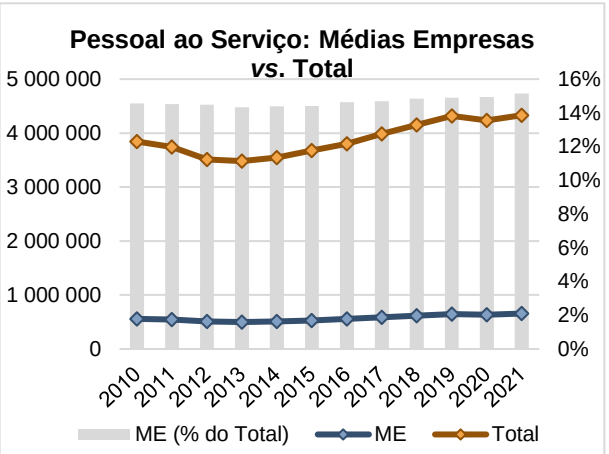


*Figura 2.7 - Pessoal ao Serviço das Empresas em Portugal: Médias Empresas vs. Total (2010-2021)*

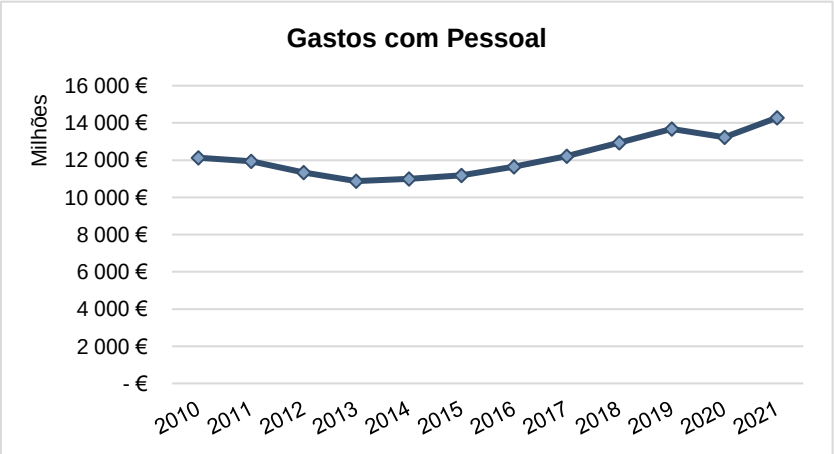


*Figura 2.8 - Gastos com Pessoal (preços constantes de 2016) das Médias Empresas em Portugal (2010-2021)*

Quanto ao pessoal ao serviço das Médias Empresas, como se pode apreciar na Figura 2.6, o número de indivíduos empregados tem aumentado desde 2013, simultaneamente com o valor de gastos em pessoal, como se pode apreciar na Figura 2.8, visto que estes indicadores estão obviamente fortemente correlacionados. Mais uma vez se verificou uma quebra no crescimento em 2020 e respetiva recuperação em 2021, relacionadas com a pandemia já mencionada.

Por outro lado, na Figura 2.7 apresenta-se a percentagem de pessoal ao serviço das Médias Empresas relativamente ao total das empresas portuguesas. Observa-se que, tal como verificado na análise do volume de negócios, esta percentagem não varia significativamente ao longo dos anos, sendo que, em média, o pessoal ao serviço das Médias Empresas representa cerca de 14,6% do total do pessoal ao serviço das empresas portuguesas.

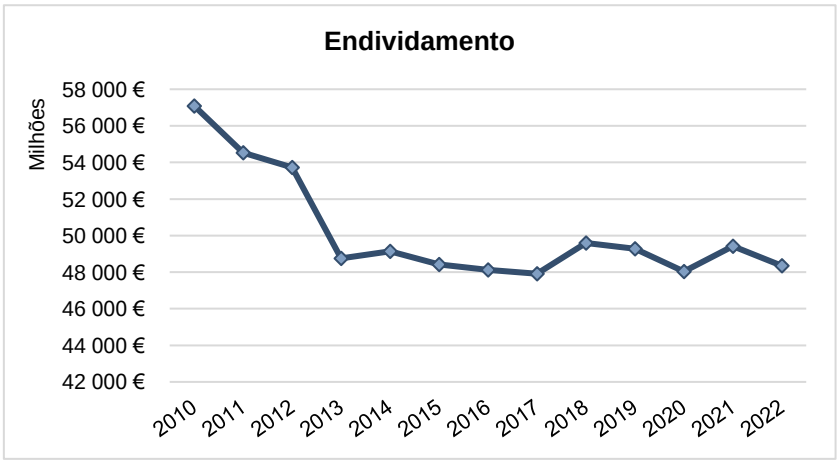


*Figura 2.9 – Endividamento (preços constantes de 2016) das Médias Empresas em Portugal (2010-2022)*

Quanto ao endividamento, na Figura 2.9, este compreende os passivos sob a forma de empréstimos, títulos de dívida (valor nominal) e créditos comerciais, sendo que no caso de empresas da administração central incluem-se ainda os certificados de aforro, certificados do Tesouro e outras responsabilidades do Tesouro. De um modo geral, o endividamento destas empresas tem vindo a diminuir significativamente, sendo que entre 2010 e 2022 houve uma redução deste indicador em cerca de 18%.

Em suma, é possível observar um claro crescimento do negócio das Médias Empresas em Portugal desde 2013 (consequência da recuperação da crise financeira em Portugal) até ao presente, com um decréscimo no ano de 2020, que reflete o início do período pandémico COVID-19, podendo-se, no entanto, observar uma imediata recuperação no ano seguinte para as Médias Empresas portuguesas.

Assim, é possível concluir que este tipo de empresas, embora represente apenas uma mínima percentagem das empresas em Portugal, se encontra em crescimento, sendo assim de esperar que muitas recorram a pedidos de crédito a bancos e que se encontrem em condições de cumprir com as suas obrigações de crédito, sendo favorável aos bancos conceder crédito a este tipo de empresas, embora não seja o seu principal negócio, como as grandes empresas. Deste modo, é de grande importância a existência de modelos que permitam aos bancos tomar a decisão de conceder, ou não, crédito a Médias Empresas.

### 2.2.2. Conceitos de Contabilidade e Análise Financeira

De seguida apresentam-se alguns conceitos de contabilidade financeira que foram utilizados neste projeto.

No momento da constituição de uma empresa é natural que os sócios depositem um capital inicial que permite o início da atividade da empresa, que se denomina como capital social. O objetivo deste capital é o de munir a empresa com meios suficientes para dar início à sua atividade económica. Com o capital social inicial e o consequente início da sua atividade económica, a empresa começa a ter resultados, estando estes geralmente ligados tanto ao consumo de recursos como à venda de bens e serviços. Assim, a soma do capital social e dos resultados obtidos até ao momento constitui o capital próprio, sendo que este representa o valor contabilístico de uma empresa num determinado momento. Note-se que a empresa pode optar por considerar parte dos resultados como bloqueados a uma posterior distribuição de resultados aos sócios, designando-se os mesmos por reservas e continuando a fazer parte do capital próprio.

Quanto aos resultados da empresa, estes podem ser divididos em vários tipos, de entre os quais se destacam os seguintes:

- Resultados Operacionais: refletem os ganhos e perdas resultantes da atividade principal da empresa, representando a capacidade do negócio principal da empresa para gerar excedentes;
- Resultados Financeiros: pretendem apurar os ganhos e perdas resultantes das decisões financeiras da empresa, abrangendo todos os custos suportados pela utilização de recursos financeiros e os proveitos resultantes de aplicações financeiras;
- Resultados Correntes: consistem na soma dos resultados operacionais e financeiros e traduzem os resultados da atividade normal da empresa, isto é, das decisões relacionadas com a exploração corrente;
- Resultados Líquidos do Exercício: consistem no apuramento do resultado líquido de cada exercício económico, correspondendo ao valor resultante de abater os custos necessários e os impostos sobre os lucros.

Para além de angariarem capital através da emissão de títulos de capital ou da realização da sua atividade económica com produção de resultados positivos, as empresas também podem obter capital a partir de terceiros, sendo este capital denominado como passivo. O passivo é cedido temporariamente por terceiros sob termos que são negociados entre ambas as partes.

Deste modo, o capital próprio e o passivo formam o total de capital que se encontra disponível para uma empresa, num determinado momento. Com esse capital, a empresa aplica-o em diferentes recursos, sendo a aplicação desse conjunto de recursos designado como o ativo da empresa.

O ativo pode ser dividido em dois grupos: o ativo corrente, que integra aqueles cujo objetivo é que se convertam direta ou indiretamente em meios líquidos, ou que são já em si meios líquidos; e o ativo fixo, que integra os ativos cujo objetivo é o de capacitar a execução da atividade económica da empresa. Com a exceção dos terrenos, imóveis e investimentos financeiros, admite-se que todos os ativos se desvalorizam ao longo do tempo, sendo que as desvalorizações de ativos intangíveis são denominadas amortizações e as desvalorizações de ativos tangíveis são designadas por depreciações.

Tem-se, assim, a regra fundamental da contabilidade:

$$Ativo\ total\ =\ Capital\ Próprio\ +\ Passivo\ total$$

Existem ainda outros termos que serão mencionados aquando da análise das características das empresas presentes na base de dados em estudo neste projeto, cujos significados se apresentam de seguida.

O volume de faturação de uma empresa corresponde ao valor total das faturas emitidas pela empresa aos seus clientes durante um determinado período. É o valor total das vendas registadas nos documentos fiscais ou comerciais emitidos pela empresa, sendo uma medida contabilística e fiscal, que está sujeita a impostos e outros encargos específicos.

O *cash-flow* de uma empresa, também designado por fluxo de caixa, corresponde ao dinheiro que entra e sai da mesma, num determinado período, distinguindo-se do lucro, uma vez que uma empresa pode ter lucros positivos e ao mesmo tempo ter os fluxos de caixa negativos por, por exemplo, não estar a receber atempadamente o reembolso dos fornecimentos aos clientes e já ter pagado as matérias-primas e os salários dos seus trabalhadores. Em suma, o *cash-flow* define-se como a diferença entre as receitas correntes e as despesas correntes de uma empresa, sendo o seu cálculo baseado no resultado líquido, a amortização e ainda algumas provisões.

As provisões de uma empresa são definidas como o "passivo de tempestividade ou quantia incerta". As provisões traduzem uma obrigação relativa a terceiros, cuja liquidação se espera que resulte numa saída de caixa de recursos da entidade que incorporem benefícios económicos, sendo as provisões diferentes de outros passivos pelo facto de a entidade desconhecer o momento em que a liquidação ocorrerá e/ou a quantia da referida obrigação.

Os custos financeiros de uma empresa correspondem aos custos associados ao financiamento da sua atividade através de recursos de terceiros, estando relacionados com juros, taxas e encargos pagos pela empresa em empréstimos, financiamentos, obrigações, entre outros.

Os acréscimos e diferimentos de uma empresa são conceitos contabilísticos associados ao momento de reconhecimento de receitas e despesas. Os acréscimos são valores pendentes de registo que ocorreram durante o período contabilístico atual, enquanto os diferimentos são receitas ou despesas que foram antecipadamente recebidas ou pagas, mas que serão registadas em períodos futuros, quando apropriadamente reconhecidas. Os acréscimos e diferimentos no passivo estão relacionados com obrigações financeiras pendentes e receitas recebidas antecipadamente, enquanto no ativo estão relacionados com ativos financeiros pendentes e despesas pagas antecipadamente.

Tal como o nome indica, os saldos médios de credores e devedores de uma empresa são medidas financeiras que representam a média dos valores que a empresa deve a fornecedores (credores) e os valores que a empresa tem a receber de clientes (devedores) ao longo de um determinado período.

Por fim, um dos tipos de descontos concedidos (ou obtidos) na atividade empresarial são os descontos comerciais. Os descontos comerciais resultam das práticas estabelecidas na atividade e que geralmente se concretizam em descontos de quantidade (como, por exemplo, descontos no preço de venda/compra). Os descontos comerciais concedidos são reduções do lucro do vendedor, sendo que os descontos comerciais obtidos reduzem o custo dos bens ou serviços adquiridos, pelo que, desta forma, quando estes bens ou serviços são incorporados na atividade, se veem os respetivos gastos reduzidos.

Sabe-se ainda que uma empresa deve realizar uma análise financeira para retirar conclusões acerca da sua situação efetiva. Esta análise consiste na produção de indicadores que traduzem a realidade económica e financeira e que pretendem traduzir os valores registados nos documentos da empresa em algo que os coloque numa base relativa, de modo a ser possível extrair conclusões e efetuar comparações entre empresas distintas. Esses indicadores são designados informalmente por rácios.

De seguida, apresentam-se alguns rácios financeiros que foram utilizados no presente projeto.

A autonomia financeira representa a proporção dos ativos que não estão caucionados por dívidas a terceiros, considerando-se assim como meios próprios. Uma autonomia financeira superior a 0,5 significa que o capital próprio representa mais de metade do ativo total, o que corresponde, de forma genérica, a uma maior independência da empresa face a credores. No entanto, nem sempre é fácil ter uma predominância tão elevada de capital próprio na empresa, uma vez que não é fácil captar capital social e que muitas vezes depende da liquidez dos investidores. Por outro lado, ainda que seja possível o capital próprio aumentar substancialmente em função de um bom desempenho económico da empresa, traduzido em resultados positivos, é natural que a empresa tenha interesse em distribuir dividendos, já que isso mantém o interesse de futuros investidores e/ou credores. Pelo contrário, e sem prejuízo do exposto anteriormente, uma autonomia financeira que apresente valores muito reduzidos significa que existe uma elevada dependência de credores e antecipa uma possível dificuldade da empresa em sobreviver financeiramente a médio/longo prazo, pelo que uma empresa nessa situação deverá rever a sua política de financiamento e/ou equacionar a angariação de capital próprio.

$$\boldsymbol{Autonomia\ Financeira} = \frac{\textbf{Capital Próprio}}{\textbf{Ativo}}$$

O rácio *debt-to-equity* relaciona o capital próprio com o passivo, correspondendo ao valor do passivo em proporção do capital próprio, sendo equivalente ao rácio anterior. Caso esta proporção se torne elevada, podendo ultrapassar a unidade, isso corresponderá a uma situação de autonomia financeira inferior a 0,5. Por outro lado, um valor reduzido do *debt-to-equity* representa um valor do passivo reduzido face ao capital próprio.

$$\boldsymbol{Debt - to - Equity} = \frac{\textbf{Passivo}}{\textbf{Capital Próprio}}$$

Os dois rácios apresentados acima avaliam a solvabilidade da empresa, que corresponde à capacidade de a empresa ser viável do ponto de vista financeiro a médio/longo prazo.

Outro tipo de avaliação a realizar na empresa relaciona-se com o seu desempenho económico, designada por análise de rendibilidade, tipicamente traduzida na proporção que representam os resultados obtidos num exercício sobre os meios que foram colocados à disposição da empresa.

A rendibilidade dos capitais próprios, também designada por *return on equity* ou ROE, avalia o resultado em função do capital próprio de que a empresa dispõe. O ROE é uma medida observada do ponto de vista dos investidores, uma vez que traduz quanto rendem os capitais próprios da empresa, medindo a rendibilidade da empresa em termos dos meios financeiros próprios da empresa.

$$\boldsymbol{ROE = \frac{\text{Resultado Líquido}}{\text{Capital Próprio}}}$$

Para além dos rácios acima apresentados, foram ainda construídos três rácios que não fazem diretamente parte da análise financeira formal das empresas, embora existam rácios financeiros oficiais muito semelhantes aos mesmos. Note-se que a opção pela criação destes rácios específicos, bem como a escolha de apenas os três rácios anteriores, se prendeu com a disponibilidade de variáveis na base de dados que serviu de suporte ao projeto.

- $\boldsymbol{\frac{Cash-Flow}{\text{Ativo}}}$
  Relaciona o *cash-flow* com o ativo da empresa. Traduz quanto rende o ativo, medindo a rendibilidade do mesmo em termos dos *cash-flows* da empresa. Em suma, este rácio mostra a proporção do ativo que resultou em *cash-flow* para a empresa.
- $\boldsymbol{\frac{Cash-Flow}{\text{Custos Financeiros}}}$
  Relaciona o *cash-flow* com os custos financeiros da empresa. Este rácio permite avaliar a capacidade da empresa de gerar dinheiro suficiente para cobrir os seus encargos financeiros. Uma relação maior do que 1 indica que o *cash-flow* é suficiente para cobrir os custos financeiros, o que é positivo, pois significa que a empresa tem uma margem de segurança para cumprir os seus compromissos financeiros. Por outro lado, uma relação inferior a 1 indica que o *cash-flow* é insuficiente para cobrir os custos financeiros, o que pode ser um sinal de alerta, pois significa que a empresa pode ter dificuldades em pagar os juros e outras obrigações financeiras.
- $\boldsymbol{\frac{\text{Passivo} - \text{Acréscimos e diferimentos (Passivo)}}{\text{Ativo} - \text{Acréscimos e diferimentos (Ativo)}}}$
  Traduz a relação entre o endividamento e os recursos próprios da empresa, ajustados pelos acréscimos e diferimentos contabilísticos, relação muitas vezes designada por alavancagem financeira. Ao calcular este rácio, obtém-se uma medida da proporção entre os recursos financeiros de terceiros e os recursos próprios da empresa. Um valor superior a 1 indica que a empresa está a financiar uma parte significativa dos seus ativos com recursos de terceiros, enquanto um valor inferior a 1 indica uma maior proporção de recursos próprios em relação aos recursos de terceiros.

# 3. Metodologias Utilizadas

O estudo realizado neste projeto foi suportado em diversas metodologias estatísticas, cuja descrição se apresenta abaixo. A metodologia principal em que assenta a construção de um modelo que permite estimar a probabilidade de incumprimento de um cliente é a regressão logística, apresentada na secção 3.5, método que permite obter um valor para essa probabilidade como função de um grupo de outros fatores, usualmente denominados por variáveis explicativas ou independentes. No entanto, um bom modelo deverá ser parcimonioso, tanto no que respeita ao número de variáveis independentes que utiliza, como também ao número de níveis das variáveis categóricas. Nesse sentido, o método de Ward (secção 3.1) em conjunção com o teste do qui-quadrado (secção 3.2) serão utilizados para reduzir o número de níveis desse tipo de variáveis. Após a preparação das potenciais variáveis a incluir no modelo, importa eliminar aquelas que podem causar problemas de multicolinearidade ou, dito de uma forma mais simples, que trazem informação que já está contida noutras variáveis. O problema da multicolinearidade pode conduzir a resultados errados nos testes de significância de algumas variáveis e, neste trabalho, esse problema foi resolvido recorrendo à correlação de Pearson, na secção 3.3, e ao método dos Fatores de Inflação da Variância, apresentado na secção 3.4.

Finalmente, o problema da seleção de variáveis pode ser resolvido através de várias metodologias cujo objetivo é o de eliminar do modelo as variáveis que não têm uma contribuição significativa para a explicação da variável resposta. Neste trabalho utilizou-se o método *stepwise* que é, talvez, o mais utilizado na escolha de modelos e que se explica na secção 3.6. Se, após a aplicação dos métodos de escolha de variáveis, o modelo contiver ainda um número muito grande de variáveis explicativas, uma possibilidade de redução desse número é através da aplicação do método das componentes principais (secção 3.7), método que foi também experimentado na tentativa de obter um modelo de regressão logística capaz de prever as empresas mais propensas ao incumprimento, modelo esse que se pretende que seja, simultaneamente, parcimonioso e com boa qualidade preditiva.

## 3.1. Método de Ward

O método de Ward foi utilizado neste projeto numa fase inicial, em que, antes de se construir qualquer modelo, se considerou importante analisar as variáveis presentes na base de dados. Na fase preliminar de análise de variáveis, mais precisamente, do conjunto das variáveis categóricas, foi identificada uma variável com um número muito elevado de categorias: a variável que representa o setor de atividade da empresa. Pretendeu-se, através deste método, identificar possíveis agrupamentos das diferentes categorias, com base na sua taxa de incumprimento, com o objetivo de reduzir o número de categorias da variável.

A análise de *clusters* é um método estatístico utilizado para organizar conjuntos de dados em grupos (ou *clusters*) com base no seu grau de homogeneidade.

Segundo Everitt et al. (2011), na classificação hierárquica, os dados não são particionados num número específico de *clusters* numa única etapa. Esta classificação consiste numa série de partições, que podem ir de um único *cluster* contendo todas as observações até *n clusters*, cada um contendo uma única observação. As técnicas de agrupamento hierárquico podem ser subdivididas em métodos aglomerativos, que se processam por uma série de fusões sucessivas das *n* observações em grupos, e métodos divisivos, que separam as *n* observações sucessivamente em grupos de menor dimensão. O

Método de Ward, também conhecido como Método da Variância Mínima, é um tipo de método hierárquico aglomerativo, descrito abaixo.

Em 1963, Ward introduziu um método no qual a fusão de dois *clusters* é baseada na dimensão de um critério da soma de quadrados dos erros. O objetivo em cada etapa é minimizar a variância dentro de cada *cluster*, $E$, dada por

$$E = \sum_{m=1}^{g} E_m,$$

onde

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} (x_{ml,k} - \bar{x}_{m,k})^2,$$

em que $\bar{x}_{m,k} = \left(\frac{1}{n_m}\right) \sum_{l=1}^{n_m} x_{ml,k}$ (média do $m$-ésimo *cluster* da $k$-ésima variável), sendo $x_{ml,k}$ o valor na $k$-ésima variável ($k = 1, \dots, p$) para a $l$-ésima observação ($l = 1, \dots, n_m$) no $m$-ésimo *cluster* ($m = 1, \dots, g$). Note-se que neste projeto será utilizado o Método de Ward aplicado a uma única variável, isto é, $p = 1$.

Em suma, em cada etapa da implementação deste método, é identificado o par de *clusters* que leva ao aumento mínimo na variância total dentro do *cluster* após a fusão.

Os resultados dos processos de análise de *clusters* são muitas vezes representados por dendrogramas, ou diagramas de árvore, que são representações gráficas do procedimento completo de *clustering*. Os nós do dendrograma representam os vários *clusters* e o comprimento dos "caules" (denominado por altura) representa a distância em que os *clusters* são agrupados.

## 3.2. Teste de Homogeneidade do Qui-Quadrado

Após aplicação do método de Ward, descrito no subcapítulo 3.1, para cada número de *clusters* selecionado com base nos resultados obtidos, foram realizados testes estatísticos de homogeneidade dentro de cada *cluster* e um teste de heterogeneidade entre os diferentes *clusters*, com o objetivo de selecionar o melhor agrupamento final para a variável que representa o setor de atividade da empresa. Ambos os testes mencionados consistem num teste de homogeneidade do qui-quadrado.

O teste do qui-quadrado é um teste não-paramétrico utilizado para três finalidades distintas: testar a qualidade do ajuste de uma distribuição observada a uma distribuição teórica esperada, testar a independência de duas variáveis categóricas ou testar a homogeneidade de grupos baseados na mesma variável categórica. Foquemo-nos no teste de homogeneidade do qui-quadrado.

Considerem-se $c$ populações independentes, $X_1, \dots, X_c$, todas com o mesmo domínio de variação particionado nas classes/categorias $A_1, \dots, A_r$. É recolhida de cada população $X_j$ uma amostra aleatória de $n_{.j}$ indivíduos e conta-se quantos deles, de acordo com as correspondentes observações de $X_j$, pertencem à categoria $A_i$, $i = 1, \dots, r$. Note-se que, em virtude de $A_1, \dots, A_r$ ser uma partição, cada indivíduo pertence a uma e uma só categoria.

Considere-se a seguinte notação:

- $n_{ij}$: número de indivíduos da população $X_j$ classificados em $A_i$, $i = 1, \dots, r$; $j = 1, \dots, c$;
- $n_{i.}$: número total de indivíduos classificados em $A_i$, $i = 1, \dots, r$, com $n_{i.} = \sum_{j=1}^{c} n_{ij}$;
- $n$: número total de indivíduos (de todas as populações) classificados, com $n = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$.

As contagens observadas podem ser organizadas numa tabela de contingência, onde cada linha está associada a uma classe $A_i$, $i = 1, \dots, r$, e cada coluna está associada a uma população $X_j$, $j = 1, \dots, c$.

As hipóteses em teste são as seguintes:

$$H_0: X_1, \dots, X_c \text{ têm distribuições homogéneas}$$

$$vs.$$

$$H_1: \text{Pelo menos uma das populações } X_1, \dots, X_c \text{ tem uma distribuição diferente das restantes.}$$

Para um indivíduo escolhido ao acaso, sejam:

- $p_{i|j}$ a probabilidade de um indivíduo pertencer à categoria $A_i$ sabendo que o indivíduo provém da população $X_j$, $i = 1, \dots, r$; $j = 1, \dots, c$;
- $p_{i.}$ a probabilidade de um indivíduo pertencer à categoria $A_i$, $i = 1, \dots, r$;
- $p_{.j}$ a probabilidade de um indivíduo provir da população $X_j$, $j = 1, \dots, c$.

Note-se que se tem $\sum_{i=1}^{r}\sum_{j=1}^{c} p_{i|j} = \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{c} p_{.j} = 1$.

De acordo com as probabilidades definidas acima, as hipóteses em teste podem ser reescritas da seguinte forma:

$$H_0: p_{i|j} = p_{i.}, \forall j = 1, \dots, c \quad vs. \quad H_1: \exists i = 1, \dots, r, \exists j = 1, \dots, c: p_{i|j} \neq p_{i.}$$

Designando $e_{i|j}$ como o número esperado de indivíduos, dos $n_{.j}$ que provêm da população $X_j$, que são classificados em $A_i$, tem-se que $e_{i|j} = n_{.j}p_{i|j}, i = 1, \dots, r; j = 1, \dots, c$ e, sob a validade da hipótese nula, $e_{i|j} = n_{.j}p_{i.}, i = 1, \dots, r; j = 1, \dots, c$.

Dado que $p_{i.}, i = 1, \dots, r$, e $p_{.j}, j = 1, \dots, c$, são desconhecidas, podem ser estimadas, como usualmente, por $\widehat{p_{i.}} = \frac{n_{i.}}{n}$ e $\widehat{p_{.j}} = \frac{n_{.j}}{n}$, respetivamente.

Assim, tem-se $\widehat{e_{i|j}} = \widehat{p_{i.}}n_{.j} = \frac{n_{i.}}{n}n_{.j} = \frac{n_{i.}n_{.j}}{n}, i = 1, \dots, r; j = 1, \dots, c$.

Deste modo, a estatística de teste utilizada para testar a homogeneidade das *c* populações em estudo é dada por:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \widehat{e_{i|j}})^2}{\widehat{e_{i|j}}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{{n_{ij}}^2}{n_{i.}n_{.j}} - 1\right),$$

que, sob a validade da hipótese nula, tem uma distribuição assintótica de um qui-quadrado com $(r-1) \times (c-1)$ graus de liberdade, isto é, $X^2|_{H_0} \sim {\chi^2}_{(r-1)\times(c-1)}$.

## 3.3.Correlação de Pearson

A correlação de Pearson, tal como as metodologias anteriormente descritas, foi utilizada também numa fase do estudo antes da criação de eventuais modelos de regressão logística. Com esta metodologia, pretendeu averiguar-se a existência de variáveis absolutamente contínuas fortemente correlacionadas entre si, com o objetivo de, posteriormente, eliminar uma das duas variáveis do par identificado.

O coeficiente de correlação de Pearson, $\rho$, mede a força da relação linear entre duas variáveis aleatórias $X$ e $Y$, calculando-se da seguinte forma:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}},$$

onde $Cov(X,Y)$ representa a covariância entre as variáveis $X$ e $Y$ e $Var(X)$ e $Var(Y)$ representam as variâncias das variáveis $X$ e $Y$, respetivamente.

Este coeficiente toma valores no intervalo $[-1;\ 1]$ e $\rho = \pm 1$ representa uma associação linear perfeita entre $X$ e $Y$, enquanto $\rho = 0$ significa que não há qualquer associação linear.

É possível estimar a força da relação linear entre $X$ e $Y$ utilizando como base uma amostra aleatória do par $(X,Y)$, nomeadamente, $(X_1,Y_1),\dots,(X_n,Y_n)$, através do cálculo do chamado coeficiente de correlação amostral de Pearson de $(X,Y)$. O coeficiente de correlação amostral de Pearson, representado por $r$, é uma medida da direção e grau com que duas variáveis quantitativas se associam linearmente. Representando a amostra bivariada observada por $(x,y) = \{(x_i,y_i)\}$, com $i = 1,\dots,n$, o coeficiente de correlação amostral de Pearson calcula-se da seguinte forma:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}},$$

onde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Deste modo, o coeficiente de correlação $r$ para o par de variáveis $(x,y)$ é o quociente entre a covariância amostral das variáveis $x$ e $y$ e o produto dos respetivos desvios-padrão:

$$r = \frac{cov(x,y)}{s_x s_y}.$$

Destacam-se as seguintes propriedades deste coeficiente:

- Tal como mencionado para o coeficiente de correlação de Pearson aplicado a duas variáveis aleatórias, o coeficiente de correlação amostral de Pearson assume valores entre -1 e 1 e, quanto maior for o valor de $r$, em módulo, maior será o grau de associação linear entre as duas variáveis;
- Um valor de $r$ positivo indica uma associação linear positiva entre as duas variáveis, isto é, quando os valores de uma das variáveis aumentam, existe tendência para que os valores da outra variável também aumentem. Por outro lado, um valor negativo de $r$ indica uma associação linear negativa entre as duas variáveis, isto é, quando os valores de uma das variáveis aumentam, existe tendência para que os valores da outra variável diminuam.

## 3.4.Fatores de Inflação da Variância (VIFs)

A análise dos fatores de inflação da variância foi uma metodologia abordada para reduzir a dimensionalidade das variáveis dos modelos a construir. Este método foi utilizado com o intuito de averiguar se seria possível eliminar variáveis com problemas de multicolinearidade que não teriam sido resolvidos através da análise inicial do coeficiente de correlação de Pearson.

Uma forma de tratar problemas de multicolinearidade entre variáveis independentes de um modelo de regressão é avaliar a qualidade dos estimadores obtidos, quando considerada em termos da grandeza das suas variâncias, sendo que esta pode ser seriamente afetada se existirem variáveis independentes que estejam relacionadas entre si. Esta eventual situação provoca um mau ajustamento do modelo no sentido em que o resultado de possíveis testes nos coeficientes de regressão pode ser enganador.

Uma maneira de verificar o grau de dependência entre cada variável independente $x_j$ e as restantes variáveis incluídas no modelo é através da examinação dos valores de $R_j^2, j = 1, \dots, p$, em que $R_j^2$ representa o valor de $R^2$ quando se faz a regressão linear múltipla de $x_j$ sobre o conjunto das restantes variáveis, ou seja, o valor de $R^2$ para o modelo

$$x_j = a_1 x_1 + \cdots + a_{j-1} x_{j-1} + a_{j+1} x_{j+1} + \cdots + a_p x_p + \eta.$$

A tolerância da variável $x_j$, designada por $TOL_j$, define-se como $TOL_j = 1 - R_j^2$. Deste modo, se $TOL_j$ se encontrar próxima do valor unitário, significa que a variável $x_j$ é independente das restantes, enquanto que, se estiver próxima do valor nulo, significa que existe uma relação aproximadamente linear entre $x_j$ e alguma das outras variáveis independentes.

O fator de inflação da variância (*variance inflation factor*), que se representa por $VIF_j$, é o inverso de $TOL_j$, isto é, $VIF_j = \frac{1}{TOL_j}$. Assim, um valor de $VIF_j$ próximo de 1 indica que não há dependência, ao passo que valores grandes indicam a presença de multicolinearidade.

Uma das maneiras mais utilizadas para esta análise é através da utilização da matriz de correlações das variáveis independentes $x_j$, $R$, uma vez que se demonstra que os elementos da diagonal principal de $R^{-1}$ correspondem exatamente aos fatores de inflação da variância. Assim, sendo $R^{-1} = [r_{ij}]$, tem-se que $r_{jj} = VIF_j$. Deste modo, é aconselhado que se retire do conjunto de variáveis independentes a variável correspondente à entrada de maior valor da diagonal principal de $R^{-1}$, se este valor for muito elevado. De seguida, após eliminação da variável selecionada, deve ser recalculada a matriz $R^{-1}$ e repetir o procedimento enquanto existirem entradas da diagonal principal com valores muito grandes, conforme o limite de tolerância escolhido.

## 3.5.Regressão Logística Múltipla

A base de todo este projeto é a regressão logística múltipla, pois é esta metodologia que permite construir um modelo de previsão da variável binária que identifica uma Média Empresa como incumpridora ou cumpridora a 12 meses. Para além da relevância deste capítulo para a construção dos modelos principais, a metodologia descrita no subcapítulo 3.5.3 foi também utilizada na análise inicial das variáveis contínuas, em que, após identificação dos pares de variáveis fortemente correlacionadas através do coeficiente de correlação de Pearson, foi realizado o teste estatístico descrito para selecionar qual das duas variáveis manter no modelo.

Quando se pretende modelar um conjunto de dados em que a variável resposta é uma variável binária, ou seja, que toma apenas os valores 0 ou 1, normalmente associada ao facto de um acontecimento ocorrer ou não, o procedimento adotado é a Regressão Logística. Seja $Y$ a variável resposta, sendo esta uma variável com distribuição de Bernoulli, isto é, tal que $P\{Y = 1\} = p$ e $P\{Y = 0\} = 1 - p$. Na regressão binária, tem-se que a probabilidade de ocorrência de um certo acontecimento (ou do respetivo complementar) varia consoante outras características associadas a esse acontecimento.

### 3.5.1. Estimação pela Máxima Verosimilhança

Comecemos por definir a função logística para o caso da Regressão Logística Simples, para depois generalizar ao caso da Regressão Logística Múltipla. Suponhamos que existem $n$ observações independentes, $Y_1, Y_2, \dots, Y_n$, de variáveis aleatórias em que cada uma possui distribuição de Bernoulli de parâmetro $p_i = p(x_i)$, em que $x_i, i = 1, \dots, n$, são observações de uma variável independente, que irão ser tratadas como um conjunto de valores constantes. Deste modo, a função massa de probabilidade de cada uma das observações é dada por:

$$P\{Y_i = y_i | x_i\} = p(x_i)^{y_i}\big(1 - p(x_i)\big)^{1-y_i},$$

em que cada $y_i$ só pode tomar o valor 0 ou 1. A função $p(x_i)$ pertence a uma família paramétrica de curvas, sendo a função logística uma das funções mais utilizadas para descrever a variação da probabilidade em termos da variação da variável independente, dependendo a mesma dos parâmetros $a$ e $b$, como descrito abaixo:

$$p(x; a, b) = \frac{e^{a+bx}}{1 + e^{a+bx}} = \frac{1}{1 + e^{-(a+bx)}}.$$

Esta função toma valores no intervalo $]0; 1[$ e é decrescente caso $b$ tome valores negativos, e crescente caso contrário.

Se quisermos generalizar ao caso de um conjunto de quaisquer $p$ variáveis explicativas, para aplicar a Regressão Logística Múltipla, é possível escrever a probabilidade de se dar um determinado acontecimento como uma função logística multivariada de um conjunto de variáveis independentes, $x_1, x_2, \dots, x_p$, que se escreve da seguinte forma:

$$p\big(x_1, \dots, x_p; b_1, \dots, b_p\big) = \frac{e^{b_1 x_1 + \cdots + b_p x_p}}{1 + e^{b_1 x_1 + \cdots + b_p x_p}} = \frac{1}{1 + e^{-(b_1 x_1 + \cdots + b_p x_p)}}.$$

Seja $b^T = [b_1 \quad b_2 \quad \dots \quad b_p]$ o vetor dos coeficientes e $x^T = [x_1 \quad x_2 \quad \dots \quad x_p]$ o vetor das variáveis.

Assim, a função logística pode ser ainda escrita como:

$$p(x;b) = \frac{e^{x^T b}}{1 + e^{x^T b}} = \frac{1}{1 + e^{-x^T b}}.$$

Defina-se a restante notação matricial. Seja $X$ a matriz de planeamento, constituída por $p$ colunas, $X = [x_1 \quad x_2 \quad \cdots \quad x_p]$, em que, em geral, na primeira todos os elementos são iguais à unidade, correspondendo ao termo constante, e as outras possuem os valores das variáveis independentes, $x_{ik}, i = 1, \dots, n, k = 1, \dots, p$. Consideremos também o vetor coluna $\Pi$ com elementos $\pi_i = p_i$, em que $p_i = p(x_1, \dots, x_p)$. Seja ainda $Y$ o vetor coluna com as observações binárias da variável dependente.

Para ajustar um modelo de regressão logística múltipla a um conjunto de vetores de observações $(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n$, deve começar-se por estimar os parâmetros $b_k, k = 1, \dots, p$, da função logística, sendo utilizado o método da máxima verosimilhança, descrito abaixo:

**1.** Cálculo da função de verosimilhança:

$$\mathrm{L}(y_1, y_2, \dots, y_n; b_1, \dots, b_p) = \prod_{i=1}^{n} p(x_{i1}, \dots, x_{ip})^{y_i} \left(1 - p(x_{i1}, \dots, x_{ip})\right)^{1-y_i} = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

**2.** Cálculo da logverosimilhança:

$$\ln\left(L(b_1, \dots, b_p)\right) = \sum_{i=1}^{n} y_i \cdot \ln(p_i) + \sum_{i=1}^{n} (1 - y_i) \cdot \ln(1 - p_i)$$

**3.** Cálculo das derivadas parciais da função logística:

$$\frac{\partial p(x_1, \dots, x_p)}{\partial b_k} = x_k \cdot p(x_1, \dots, x_p) \cdot \left(1 - p(x_1, \dots, x_p)\right) = x_k \cdot p(X) \cdot \left(1 - p(X)\right)$$

**4.** Cálculo das equações normais do modelo:

$$\frac{\partial \ln\left(L(b_1, \dots, b_p)\right)}{\partial b_k} = \frac{\partial \ln(L)}{\partial b_k} = \sum_{i=1}^{n} y_i \cdot \frac{\partial p_i / \partial b_k}{p_i} - \sum_{i=1}^{n} (1 - y_i) \cdot \frac{\partial p_i / \partial b_k}{1 - p_i} =$$

$$= \sum_{i=1}^{n} x_{ik} \cdot (y_i - p_i) = 0, k = 1, \dots, p.$$

Utilizando a notação matricial:

$$\frac{\partial \ln(L)}{\partial b} = X'(Y - \Pi) = 0.$$

Estas equações normais não possuem solução analítica, pelo que é necessário recorrer a métodos numéricos como, por exemplo, o método de Newton-Raphson. Este é um método iterativo em que cada iteração é obtida a partir da anterior através da expressão

$$b^{(j+1)} = b^{(j)} - H_j^{-1} \left.\frac{\partial \ln(L)}{\partial b}\right|_{b^{(j)}},$$

em que $H_j^{-1}$ representa a inversa da matriz Hessiana, que é a matriz das segundas derivadas da logverosimilhança calculada em $b^{(j)}$ e em que as derivadas da logverosimilhança são também calculadas nesse ponto. A matriz das segundas derivadas tem o elemento genérico:

$$\frac{\partial^2 \ln(L)}{\partial b_l \partial b_k} = -\sum_{i=1}^{n} x_{ik} \cdot x_{il} \cdot p_i \cdot (1 - p_i)\,, k = 1, \dots, p \quad \text{e} \quad l = 1, \dots, p.$$

Assim, se considerarmos uma matriz diagonal $V$, em que os elementos da diagonal principal são da forma $p_i \cdot (1 - p_i), i = 1, \dots, n$, a matriz das segundas derivadas da logverosimilhança de um modelo de regressão logística é dada por $H = -X'VX$. Note-se que $-H = X'VX$ é a matriz de informação de Fisher, geralmente designada por $I(b)$. Deste modo, tem-se:

$$b^{(j+1)} = b^{(j)} + \left(X'V_jX\right)^{-1} \frac{\partial \ln(L)}{\partial b}\bigg|_{b^{(j)}} = b^{(j)} + I(b)^{-1} \frac{\partial \ln(L)}{\partial b}\bigg|_{b^{(j)}} = b^{(j)} + \left(X'V_jX\right)^{-1} X'\left(Y - \Pi_j\right),$$

em que o índice $j$ na matriz $V_j$ indica que os elementos na diagonal principal são calculados tomando como parâmetros do modelo os obtidos na iteração $j$.

Os valores iniciais para os coeficientes podem ser tomados, por exemplo, considerando $p_i$ constante ou, equivalentemente, $b_2^{(0)} = b_3^{(0)} = \cdots = b_p^{(0)} = 0$ *e* $b_1^{(0)}$ tal que $p_i = p = \bar{y}$:

$$\bar{y} = \frac{e^{b_1^{(0)}}}{1 + e^{b_1^{(0)}}} \Leftrightarrow b_1^{(0)} = \ln\left(\frac{\bar{y}}{1 - \bar{y}}\right).$$

O processo termina quando duas iterações consecutivas produzirem valores suficientemente próximos, conforme critério de paragem definido à partida, através, por exemplo, do valor de tolerância $\delta$:

$$\left\|b^{(j+1)} - b^{(j)}\right\| < \delta.$$

É possível demonstrar que as equações normais de um modelo de regressão logística têm solução única e que essa solução corresponde a um máximo da verosimilhança. No entanto, o método iterativo descrito pode não convergir quando a matriz *H* tem determinante próximo de zero, o que pode acontecer devido a uma matriz de planeamento mal delineada. Assim, se, por exemplo, as variáveis independentes forem variáveis indicatrizes e se existir uma ou mais com o valor unitário numa proporção muito pequena relativamente à dimensão da amostra, o algoritmo poderá não convergir.

### 3.5.2. Inferência Estatística nos Parâmetros

Depois de estimados os coeficientes da regressão logística, importa testar a significância das variáveis incluídas no modelo.

Segundo Fahrmeir e Kaufmann (1985), em determinadas condições que envolvem a matriz de informação de Fisher *I*(*b*) e a matriz de planeamento X, tem-se que

i) $\hat{b} \xrightarrow{P} b$;

ii) $I^{1/2}(b)\left(\hat{b} - b\right) \xrightarrow{L} N(0, I)$,

em que $I^{1/2}(b)$ é uma matriz tal que, multiplicada por ela própria, é a matriz de informação de Fisher e $\hat{b}$ é o vetor dos estimadores de máxima verosimilhança do vetor de coeficientes $b$.

Posto de outra forma, os estimadores de máxima verosimilhança do vetor $\hat{b}$ são consistentes e com distribuição assintoticamente normal.

Com esta informação, é possível determinar um teste aproximado para as seguintes hipóteses:

$$H_0: b_j = 0 \qquad vs. \qquad H_1: b_j \neq 0.$$

Considerando $\hat{\sigma}^2(\widehat{b_j})$ o estimador da variância de $b_j$, isto é, o $j$-ésimo elemento na diagonal da matriz $I(b)^{-1}$, um teste para esta hipótese é dado pela região de rejeição

$$|Z| = \frac{|\widehat{b_j}|}{\hat{\sigma}(\widehat{b_j})} > q_{1-\alpha/2},$$

em que $q_{1-\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição normal padrão. Note-se que se pode substituir o desvio-padrão do estimador do coeficiente por um seu estimador porque este é consistente. Assim, qualquer função contínua do estimador, como é o caso do seu desvio-padrão, é também consistente e, no limite, não se altera a distribuição de probabilidade da estatística de teste, mantendo-se normal padrão.

### 3.5.3. Ajustamento e Escolha do Modelo

Para avaliar a qualidade do ajustamento de um modelo de regressão logística, é possível aplicar um teste de razão de verosimilhanças generalizado às hipóteses:

$$H_0: b_2 = b_3 = \cdots = b_p = 0 \qquad vs. \qquad H_1: \exists j, j = 2, \dots, p: b_j \neq 0.$$

Note-se que o termo $b_1$ corresponde ao termo constante e, por isso, não está incluído no teste. Este teste avalia sobre a necessidade de ajustar um modelo de regressão logística e é baseado na seguinte estatística:

$$\lambda = \frac{L_0}{L_1} \leq k,$$

em que $L_0$ é a função de verosimilhança calculada nos estimadores da máxima verosimilhança sob a validade da hipótese nula, ou seja, $L_0 = L(y_1, \dots, y_n; b_1)$, e $L_1$ é a verosimilhança calculada nos estimadores de máxima verosimilhança sem qualquer restrição, isto é, $L_1 = L(y_1, \dots, y_n; b_2, \dots, b_p)$. Pelo teorema de Wilks, sabe-se que $-2\ln(\lambda)$ tem distribuição assintótica qui-quadrado com $p-1$ graus de liberdade. Note-se que $\lambda = \frac{L_0}{L_1}$ é geralmente designado por *likelihood ratio* e, quanto maior for o valor da estatística $-2\ln(\lambda)$, melhor é o modelo.

Outra forma de avaliar a qualidade do ajustamento é através da comparação dos valores previstos com os valores observados. Pode definir-se um ponto de corte específico para a probabilidade prevista pelo modelo, em que, caso a probabilidade obtida pelo modelo seja inferior ao valor do ponto de corte, se considera que o valor previsto pelo modelo é 0 e, caso contrário, é 1.

De um modo geral costuma-se chamar valores positivos aos valores (observados ou previstos) iguais à unidade e valores negativos àqueles que são nulos. Os valores positivos que são previstos como tal são denominados por verdadeiros positivos (VP) e os que são previstos erradamente como negativos são denominados como falsos negativos (FN). De forma análoga, os verdadeiros valores negativos podem ser previstos corretamente como sendo negativos, sendo denominados como verdadeiros negativos (VN), ou podem ser previstos como positivos e, nesse caso, denominam-se por falsos positivos (FP). Uma vez classificados todos os pares de valores observados/previstos, os resultados são apresentados numa Matriz de Confusão, que apresenta os totais de pares em cada uma das categorias tal como se mostra na Tabela 3.1.

*Tabela 3.1 - Matriz de Confusão*

| | **Observados** | | |
|---|---|---|---|
| **Previstos** | Positivos | Negativos | **Total** |
| Positivos | VP | FP | VP+FP |
| Negativos | FN | VN | FN+VN |
| **Total** | VP+FN | FP+VN | Nº Observações |

Pretende-se que os valores de FP e de FN sejam os menores possíveis, podendo estes valores ser alterados consoante alteração do ponto de corte escolhido. No entanto, através da modificação do ponto de corte, não é possível diminuir o número de falsos positivos sem aumentar os falsos negativos, nem inversamente.

Para avaliar a qualidade do modelo, isto é, a sua percentagem de acerto, definem-se as seguintes probabilidades:

1) **Sensibilidade (S):** probabilidade de uma observação ser classificada como positiva sendo efetivamente positiva:

$$S = \frac{VP}{VP + FN}$$

2) **Especificidade (E):** probabilidade de uma observação ser classificada como negativa sendo efetivamente negativa:

$$E = \frac{VN}{VN + FP}$$

Sendo, regra geral, o valor positivo uma ocorrência que se quer evitar, deve procurar-se manter a proporção de falsos positivos $(1 - E)$ baixa, o que, se o modelo for bom, não terá como consequência uma proporção exagerada de falsos negativos $(1 - S)$.

Para avaliar o modelo e procurar o melhor ponto de corte, representa-se num gráfico a sensibilidade no eixo vertical contra $1 - E$ no eixo horizontal, resultando numa curva a que se chama curva ROC (*Receiving Operating Characteristic*). Num bom modelo, esta curva deve crescer rapidamente para o valor unitário, à medida que o ponto de corte cresce, afastando-se da diagonal, que corresponde a um modelo em que a previsão é feita completamente ao acaso. Deste modo, uma curva ROC representa, para cada valor do ponto de corte, a percentagem de verdadeiros positivos contra a percentagem de falsos positivos, apresentando-se um exemplo deste tipo de curva na Figura 3.1.

A área abaixo da curva ROC, comummente designada como AUC (*Area Under the Curve*), que varia entre zero e um, fornece uma medida da capacidade discriminatória do modelo entre os dois possíveis valores da variável resposta, sendo que uma área de 0,5 corresponde ao evento completamente aleatório, não sugerindo qualquer discriminação. Deste modo, quanto maior o valor de AUC, melhor o poder discriminatório do modelo, traduzindo a capacidade do mesmo de efetuar previsões corretas e distinguir corretamente os eventos positivos e negativos.

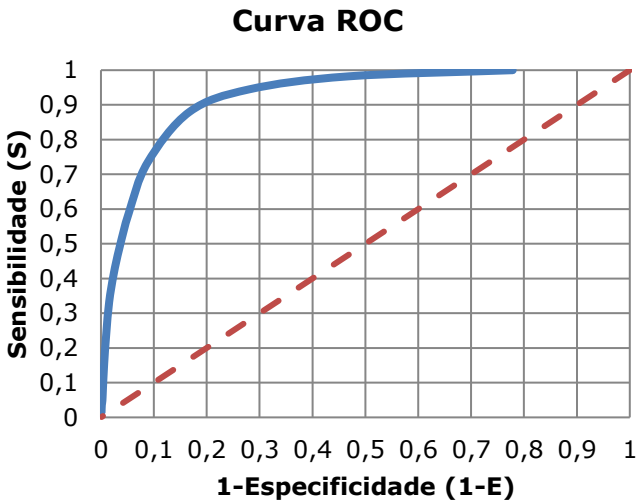


*Figura 3.1 - Curva ROC*

Outro método estatístico utilizado para avaliar e comparar modelos em análise de regressão logística é através do AIC (*Akaike Information Criterion*). Este indicador é calculado com base na função de verosimilhança do modelo e na penalização pelo número de parâmetros estimados, sendo dado por:

$$AIC = -2\ln\left(\frac{L_0}{L_1}\right) + 2k = -2\ln(\lambda) + 2k,$$

onde $k$ corresponde ao número de parâmetros estimados no modelo.

O objetivo deste critério é encontrar um equilíbrio entre o ajuste do modelo aos dados e a sua complexidade. O AIC favorece modelos que se ajustam bem aos dados, mas possuem menos parâmetros, evitando assim o sobreajuste. Ao comparar diferentes modelos, aquele com o valor mais baixo de AIC é considerado o melhor, no entanto, este não fornece uma medida absoluta de ajuste, mas sim uma medida relativa entre os modelos considerados.

## 3.6.Método de Seleção *Stepwise*

Com a exceção do modelo inicial completo apresentado neste projeto, cujo objetivo foi sempre o de fornecer um enquadramento e não o de ser o modelo final, os modelos relevantes construídos neste estudo são o resultado da aplicação do método de seleção *stepwise*, que se considerou o método de seleção mais completo para contribuir para o objetivo de obter o modelo mais parcimonioso possível.

De um modo geral, pretende-se um modelo de regressão parcimonioso, isto é, incluindo o mínimo possível de variáveis e mantendo a qualidade do ajustamento. Quanto maior for o número de parâmetros a estimar, maior será a variância dos estimadores do modelo de regressão e, por isso, deve evitar-se a inclusão de variáveis que não tenham um peso significativo na explicação da variável dependente. Uma das formas de reduzir o número de variáveis independentes incluídas num modelo de regressão, e dado que a utilização de muitas variáveis aumenta a possibilidade da existência de multicolinearidade, é através de métodos de seleção, nomeadamente, o Método de Seleção *Stepwise*.

Um dos métodos de seleção de variáveis que existem é o método da seleção progressiva, que tem início com 0 variáveis no modelo e inclui, sucessivamente, aquelas que provocam um maior aumento na

qualidade do ajustamento, terminando o procedimento quando qualquer variável não incluída no modelo não é significativa. Por sua vez, o Método de Seleção *Stepwise* é um método de seleção progressiva ao qual se junta, após a inclusão de uma nova variável, um passo adicional em que se testa a significância de todas as variáveis incluídas no modelo e se retiram aquelas que não são significativas. Descrevem-se abaixo os vários passos deste tipo de método de seleção de variáveis para um modelo de regressão logística:

**1)** Começar com 0 variáveis no modelo ($p = 1$);
**2)** Dado que o modelo atual tem $p$ variáveis, considerar cada uma das restantes variáveis e escolher aquela que provoca maior aumento na qualidade do ajustamento, selecionando a que apresenta um maior valor na variável $-2\ln(\lambda) = -2\ln\left(\frac{L_0}{L_1}\right)$;
**3)** Testar se o aumento na qualidade é significativo, realizando o teste Z apresentado no capítulo 3.5.2.;
**4)** Se o resultado do teste em 3 for positivo, juntar a variável ao modelo. Se não, ir para o 6º passo;
**5)** Depois de incluída a nova variável, testar a significância das restantes variáveis já incluídas no modelo, através do mesmo teste, e retirar aquelas que não forem significativas. Ir para o passo 2;
**6)** Terminar com o modelo atual com $p$ variáveis.

Obtém-se assim, de um modo geral, um modelo com um menor número de variáveis, tornando o modelo mais interpretável, mantendo ou melhorando a sua capacidade preditiva.

## 3.7. Análise de Componentes Principais

Tal como o método de seleção *stepwise*, a análise de componentes principais neste projeto foi aplicada com o objetivo de obter um modelo mais parcimonioso, com o menor número de variáveis possíveis. Dada a existência de um número de variáveis absolutamente contínuas bastante elevado no segundo modelo construído, mesmo após aplicação do método de seleção mencionado, optou-se por realizar esta análise e substituir as variáveis contínuas pelo menor número possível de componentes principais obtidas que explicassem suficientemente as variáveis originais.

Segundo Varella (2008), a análise de componentes principais é um método estatístico cujo objetivo é reduzir a dimensionalidade das variáveis de um conjunto de dados com a menor perda de informação possível. Esta técnica baseia-se na transformação de um conjunto de variáveis originais noutro conjunto de variáveis com a mesma dimensão, designadas como componentes principais. Cada componente principal é uma combinação linear de todas as variáveis originais, sendo independentes entre si. O objetivo é redistribuir a variação observada nos eixos originais de modo a obter um conjunto de eixos ortogonais não correlacionados.

Considere-se a situação em que existe um conjunto de dados com $n$ observações distribuídas por $p$ variáveis, $X_1, \dots, X_p$. A matriz de dados, $X$, de dimensão $n \times p$, é dada por

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

O objetivo da análise de componentes principais é transformar a estrutura complicada de interdependência entre as variáveis da matriz de dados, representada pelas variáveis $X_1, \dots, X_p$, numa outra estrutura representada pelas variáveis $Y_1, \dots, Y_p$ não correlacionadas e com variâncias ordenadas, de modo a que seja possível comparar as diferentes observações da amostra utilizando apenas as variáveis $Y_j, j \in \{1, \dots, p\}$, que apresentam maior variância. Isto pode ser obtido com o auxílio da matriz de correlações $R$ ou da matriz de covariância $S$, tendo-se:

$$S = \begin{bmatrix} \widehat{Var}(X_1) & \widehat{Cov}(X_1, X_2) & \cdots & \widehat{Cov}(X_1, X_p) \\ \widehat{Cov}(X_2, X_1) & \widehat{Var}(X_2) & \cdots & \widehat{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{Cov}(X_p, X_1) & \widehat{Cov}(X_p, X_2) & \cdots & \widehat{Var}(X_p) \end{bmatrix}.$$

Note-se que a matriz supra tem dimensão $p \times p$ e é simétrica.

Como as características das diferentes variáveis são muitas vezes observadas em diferentes unidades de medida, é comum e conveniente *standardizar* (ou padronizar) as variáveis $X_j, j = 1, \dots, p$, sendo esta transformação feita normalmente de modo a obter uma variável com valor médio zero e variância unitária, obtendo-se assim a matriz *standardizada* $Z$, $Z = \{z_{ij}\},\ i \in \{1, \dots, n\}, j \in \{1, \dots, p\}$, onde

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s(x_j)}.$$

Note-se que $\overline{x_j}$ e $s(x_j)$ são a estimativa da média e do desvio-padrão da variável $X_j$, respetivamente:

$$\overline{x_j} = \hat{\overline{X}}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \qquad \text{e} \qquad s(x_j) = \sqrt{\widehat{Var}(X_j)} = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2}{n-1}.$$

Note-se ainda que na análise de componentes principais utilizada neste projeto foi utilizada a normalização das variáveis para que as mesmas tivessem valor médio nulo e desvio-padrão unitário.

A matriz $Z$ é equivalente à matriz de correlação da matriz de dados, $R$. A escolha entre a utilização da matriz de covariância ou da matriz de correlação depende do contexto da análise, das unidades das variáveis e das relações entre elas. Se se pretender ter em consideração as unidades originais das variáveis ou as diferenças nas variâncias (ou se os dados já estiverem na mesma escala), deve ser utilizada a matriz de covariância, onde os dados não são normalizados. Por outro lado, se se pretender eliminar a influência das diferentes escalas e unidades para se focar apenas na estrutura de correlações, deve ser utilizada a matriz de correlação, que mede a relação entre as variáveis após remover o efeito das diferentes escalas e unidades. Na análise realizada neste projeto, optou-se por utilizar a matriz de correlações.

As componentes principais são determinadas através da resolução da equação característica da matriz $S$ ou $R$, isto é, $\det(R - \lambda I) = 0$, ou seja, através da determinação dos vários valores próprios, $\lambda_1, \dots, \lambda_p$, e correspondentes vetores próprios, $v_1, \dots, v_p$, da matriz em questão. Note-se que para cada valor próprio $\lambda_i$ existe um e só um vetor próprio $v_i$, uma vez que cada vetor próprio tem norma unitária. Os diferentes vetores próprios são ainda ortogonais entre si. O $i$-ésimo vetor próprio é dado por:

$$v_i = \begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{ip} \end{bmatrix}.$$

Sendo $v_i$ o vetor próprio correspondente ao valor próprio $\lambda_i$, tem-se que a $i$-ésima componente principal, $Y_i$, é dada por:

$$Y_i = v_{i1}X_1 + v_{i2}X_2 + \cdots + v_{ip}X_p.$$

Destacam-se algumas propriedades das componentes principais:

**1)** A variância da componente principal $Y_i$ é dada pelo valor próprio $\lambda_i$:
$$\widehat{Var}(Y_i) = \lambda_i, i = 1, \dots, p;$$

**2)** As variâncias das componentes principais apresentam-se por ordem decrescente:
$$\widehat{Var}(Y_1) > \widehat{Var}(Y_2) > \cdots > \widehat{Var}(Y_p);$$

**3)** O total das variâncias das variáveis iniciais é igual ao somatório dos valores próprios, sendo ainda igual ao total das variâncias das componentes principais:
$$\sum \widehat{Var}(X_i) = \sum \lambda_i = \sum \widehat{Var}(Y_i), i = 1, \dots, p;$$

**4)** As componentes principais não estão correlacionadas entre si:
$$\widehat{Cov}(Y_i, Y_j) = 0, i = 1, \dots, p, j = 1, \dots, p, i \neq j.$$

A contribuição $C_i$ de cada componente principal $Y_i$, expressa em percentagem, representa a proporção da variância total que é explicada pela componente principal $Y_i$, calculada da seguinte forma:

$$C_i = \frac{\widehat{Var}(Y_i)}{\sum_{i=1}^{p} \widehat{Var}(Y_i)} \times 100 = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100.$$

A soma das contribuições das várias componentes principais (por ordem) permite escolher quantas componentes serão utilizadas na análise, de modo a que sejam suficientes para explicar os dados, explicando a maior parte da variância dos mesmos, não existindo um método exato para tomar esta decisão. É comum, para aplicações em diferentes áreas, que o número de componentes utilizadas seja aquele que acumula entre 70% e 90% da proporção da variância total. No entanto, pode considerar-se qualquer limite de tolerância à preferência da pessoa a realizar a análise.

Para interpretar cada componente obtida, é possível analisar o grau de influência de cada variável original (ou *standardizada*, conforme o método escolhido) em cada componente principal. Se se quiser analisar a componente principal $Y_i, i = 1, \dots, p$, um dos possíveis métodos de interpretação da mesma é a análise da correlação entre esta e as várias variáveis $X_j, j = 1, \dots, p$, dada por:

$$Corr(X_j, Y_i) = r_{X_j, Y_i} = v_{1j} \cdot \frac{\sqrt{\widehat{Var}(Y_i)}}{\sqrt{\widehat{Var}(X_j)}} = \sqrt{\lambda_i} \cdot \frac{v_{ij}}{\sqrt{\widehat{Var}(X_j)}},$$

que, no caso da normalização das variáveis, se simplifica ainda em:

$$Corr(X_j, Y_i) = \sqrt{\lambda_i} \cdot v_{ij}.$$

Se o objetivo for comparar a influência de $X_1, X_2, \dots, X_p$ sobre $Y_i$, pode também analisar-se o peso de cada variável sobre a respetiva componente principal, dado por:

$$w_j = \frac{v_{ij}}{\sqrt{\widehat{Var}(X_j)}},$$

sendo $w_j$ o peso de $X_j$. Se se der o caso da normalização das variáveis, o peso de cada variável traduz-se simplesmente pelo próprio coeficiente, $v_{ij}$, da variável $X_j$ na componente principal $Y_i$, sendo imediata a comparação entre as várias variáveis em cada componente principal.

# 4. Caso de Estudo

## 4.1. Análise da Base de Dados

O objetivo deste projeto é a construção de um modelo de previsão do incumprimento de Médias Empresas. Para tal, foi recolhida uma base de dados com informações relativas a Médias Empresas com crédito num determinado banco português. A base de dados que foi utilizada para a realização deste estudo tem 30.000 observações e um conjunto de variáveis relacionadas com o comportamento e atividade da empresa, sendo o objetivo averiguar se estas variáveis são suficientes para explicar o incumprimento das empresas em questão e, em caso afirmativo, qual o conjunto de variáveis mais adequado. Por questões de confidencialidade, foi recolhida uma amostra aleatória com uma percentagem de *defaults* de 25%.

Na base de dados fornecida pelo banco, todas as observações dizem respeito ao estado de um contrato de uma Média Empresa num determinado momento em que a empresa em questão não se encontra em incumprimento do seu contrato de crédito, sendo que o banco considera que uma Média Empresa engloba os clientes empresariais com um volume de faturação entre 1,25 e 50 milhões de euros.

### 4.1.1. Análise e Tratamento das Variáveis

A variável resposta, daqui adiante denominada como "Def12m", que se pretende avaliar neste estudo é uma variável binária, que toma o valor 1 se a empresa tiver entrado em incumprimento nos 12 meses seguintes à data da observação, tomando o valor 0 caso contrário. Ao observar-se um contrato que não se encontra em *default* no momento da observação e ao analisar se entrou ou não em *default* no ano seguinte, isto permite construir um modelo que permite ao banco perceber se, ainda que o contrato se encontre "saudável" no momento presente, este é provável ou não de entrar em incumprimento nos 12 meses seguintes.

As restantes variáveis presentes nesta base de dados dividem-se, de um modo geral, em dois tipos:

- Variáveis Quantitativas, sendo estas, na sua maioria, variáveis relacionadas com a *performance* financeira da empresa, como por exemplo, valores de certas rúbricas do balanço da empresa;
- Variáveis Qualitativas, sendo estas relacionadas com o conhecimento que a área comercial do banco adquiriu sobre a empresa, sendo que, por vezes, nem sempre foi possível recolher informação relativamente a todos os temas avaliados, sendo atribuída a estes casos a observação "Desconhecido".

Para além destas, possui-se ainda uma variável binária que indica se o contrato em questão se encontra em moratória[1] no banco no momento da observação. Esta variável pode eventualmente ser útil no sentido em que um contrato em moratória pode indicar possíveis dificuldades da empresa em cumprir com as suas obrigações de crédito, podendo ser indicador de que a empresa entrará em incumprimento no futuro, após desmarcação da moratória. Note-se, porém, que é possível que esta variável também não tenha importância, já que também é possível que uma empresa tenha aderido a moratória sem estar em dificuldades financeiras e apenas porque tinha acesso a essa possibilidade.

[1] Uma moratória diz respeito ao adiamento do prazo do pagamento de uma dívida, concedido pelo credor ao devedor.

#### 4.1.1.1. Variáveis Qualitativas

As variáveis qualitativas disponíveis na base de dados são como se descreve de seguida, sendo também apresentadas análises descritivas em que se relaciona a distribuição das várias categorias da variável com a variável resposta do modelo a criar.

1. **Setor de Atividade da Empresa**
   A empresa em análise pode pertencer a um dos seguintes setores de atividade:
   - Indústria Transformadora;
   - Indústria Têxtil;
   - Transportes;
   - Comércio Grosso Alimentar;
   - Serviços;
   - Comércio de Retalho;
   - Indústria de Materiais;
   - Comércio Grosso de Materiais;
   - Indústria Química;
   - Comércio de Automóveis;
   - Comércio Grosso de Equipamentos;
   - Indústria Alimentar;
   - Indústria Tecnológica;
   - Agricultura;
   - Extração;
   - *Utilities*;
   - Turismo de Hotelaria;
   - Turismo de Restauração;
   - Atividades Financeiras;
   - *Holdings*;
   - Estado;
   - Associações.

   Na Tabela 4.1 apresentam-se os setores de atividade ordenados por ordem crescente da respetiva taxa de incumprimento na amostra.

*Tabela 4.1 - Distribuição da variável "Setor de Atividade" pelas respetivas categorias*

| Setor de Atividade | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Atividades Financeiras** | 41 | 0 | 41 | 0,0% |
| **Associações** | 284 | 40 | 324 | 12,3% |
| **Comércio de Retalho** | 2 780 | 580 | 3 360 | 17,3% |
| **Estado** | 81 | 17 | 98 | 17,3% |
| **Turismo de Hotelaria** | 453 | 104 | 557 | 18,7% |
| **Comércio Grosso de Equipamentos** | 1 206 | 291 | 1 497 | 19,4% |
| **Indústria Química** | 857 | 227 | 1 084 | 20,9% |
| **Serviços** | 3 665 | 1 122 | 4 787 | 23,4% |
| **Comércio Grosso de Materiais** | 2 103 | 651 | 2 754 | 23,6% |
| **Comércio de Automóveis** | 969 | 301 | 1 270 | 23,7% |
| **Indústria Transformadora** | 465 | 147 | 612 | 24,0% |
| **Indústria de Materiais** | 2 368 | 766 | 3 134 | 24,4% |
| ***Utilities*** | 122 | 40 | 162 | 24,7% |
| **Indústria Tecnológica** | 689 | 253 | 942 | 26,9% |
| **Turismo de Restauração** | 300 | 114 | 414 | 27,5% |
| **Indústria Alimentar** | 924 | 361 | 1 285 | 28,1% |
| **Comércio Grosso Alimentar** | 1 555 | 655 | 2 210 | 29,6% |
| **Agricultura** | 550 | 243 | 793 | 30,6% |
| **Transportes** | 1 135 | 554 | 1 689 | 32,8% |
| **Indústria Têxtil** | 1 585 | 788 | 2 373 | 33,2% |
| ***Holdings*** | 234 | 135 | 369 | 36,6% |
| **Extração** | 134 | 111 | 245 | 45,3% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

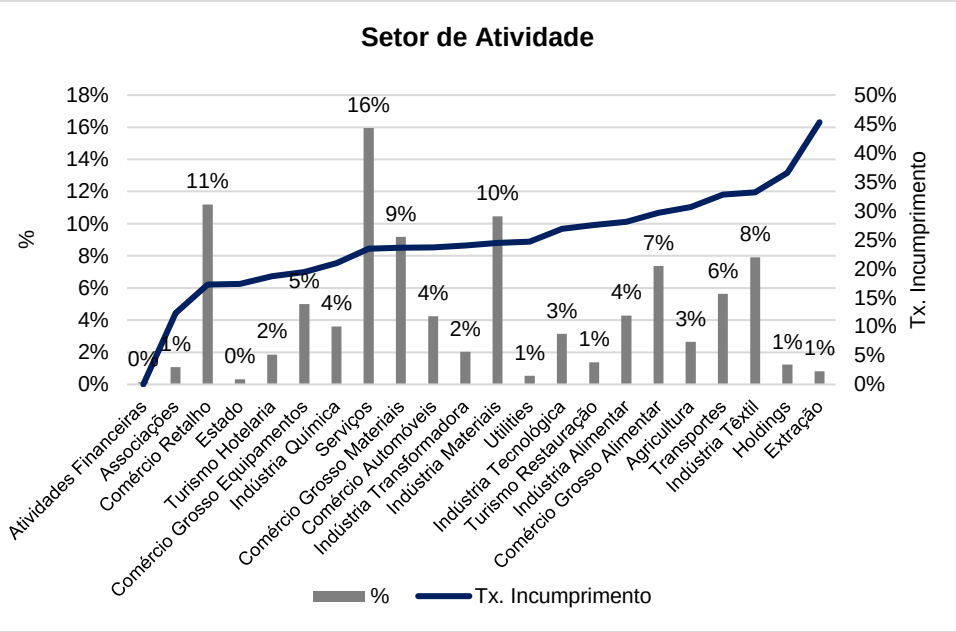


*Figura 4.1 - Setor de Atividade: Distribuição e Taxas de Incumprimento*

Esta variável está dividida em 22 categorias, algumas com taxas de incumprimento bastante próximas, pelo que se considerou importante perceber se é possível agrupá-la num menor número de *clusters*. Para tal, procedeu-se à aplicação do Método de Ward com base na taxa de incumprimento de cada setor.

Com o auxílio da aplicação SAS Enterprise Guide 9.4 PRD, foi aplicado o Método de Ward a esta variável. Para visualização gráfica, recorreu-se ao RStudio, para representação do resultado obtido com o primeiro *software* mencionado.

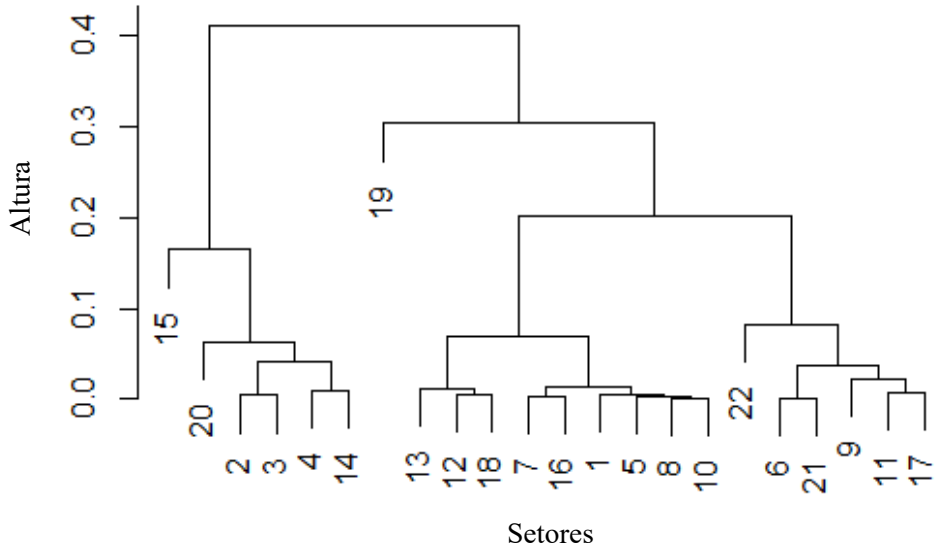


*Figura 4.2 – Dendrograma resultante da aplicação do Método de Ward à variável "Setor de Atividade"*

Na Tabela 4.2 apresenta-se a correspondência entre o identificador numérico de cada setor no dendrograma exposto na Figura 4.2 e o descritivo do respetivo setor.

*Tabela 4.2 - Correspondência entre o identificador numérico de cada setor de atividade e o seu respetivo descritivo*

| ID Setor | Setor | ID Setor | Setor |
|---|---|---|---|
| 1 | Indústria Transformadora | 12 | Indústria Alimentar |
| 2 | Indústria Têxtil | 13 | Indústria Tecnológica |
| 3 | Transportes | 14 | Agricultura |
| 4 | Comércio Grosso Alimentar | 15 | Extração |
| 5 | Serviços | 16 | *Utilities* |
| 6 | Comércio de Retalho | 17 | Turismo de Hotelaria |
| 7 | Indústria de Materiais | 18 | Turismo de Restauração |
| 8 | Comércio Grosso de Materiais | 19 | Atividades Financeiras |
| 9 | Indústria Química | 20 | *Holdings* |
| 10 | Comércio de Automóveis | 21 | Estado |
| 11 | Comércio Grosso de Equipamentos | 22 | Associações |

Como o objetivo era reduzir o número de categorias apresentadas nesta variável, foram realizados testes para diferentes números de *clusters* escolhidos para o novo agrupamento da variável. Tendo a variável original 22 categorias, optou-se por começar com um mínimo de 5 *clusters*, uma vez que se considerou, de forma qualitativa, que um número de *clusters* inferior a este valor seria demasiado reduzido para transmitir a informação contida nesta variável. Para cada número de *clusters* selecionado, foram

realizados testes estatísticos de homogeneidade dentro de cada *cluster* e um teste de heterogeneidade entre os diferentes *clusters*.

Quanto ao teste de homogeneidade, começou por realizar-se tantos testes de homogeneidade do qui-quadrado quanto o número de *clusters* selecionado em cada iteração. As hipóteses a testar, para cada grupo de setores, foram as seguintes:

$$H_0\text{: Os diferentes setores incluídos no } cluster \text{ têm distribuições homogéneas}$$

$$vs.$$

$$H_1\text{: Pelo menos um dos setores incluído no } cluster \text{ tem uma distribuição diferente dos restantes}$$

O objetivo será não rejeitar a hipótese nula, o que nos garante que os diferentes setores incluídos no *cluster* têm distribuições homogéneas, como pretendido.

Após este teste, realizou-se ainda um teste mais rigoroso, com o objetivo de testar a homogeneidade com um único teste para todos os *clusters*. As hipóteses de teste mantêm-se, sendo, no entanto, para os vários *clusters* em simultâneo. Sabemos que o somatório de variáveis independentes com distribuição qui-quadrado corresponde a uma variável com distribuição qui-quadrado em que os respetivos graus de liberdade correspondem à soma dos graus de liberdade das variáveis incluídas no somatório. Assim, a estatística de teste deste segundo teste corresponde à soma das várias estatísticas de teste dos testes de homogeneidade individuais realizados anteriormente e já mencionados, tendo esta estatística de teste uma distribuição qui-quadrado com tantos graus de liberdade quanto a soma dos graus de liberdade das estatísticas de teste dos testes individuais por *cluster*. O objetivo mantém-se o de não rejeitar a hipótese nula, uma vez que, na sua essência, mantêm-se também as hipóteses a testar, sendo que não rejeitar a hipótese nula nos indica que todos os *clusters* considerados incluem setores homogéneos entre si, como é desejado.

Quanto ao teste de heterogeneidade, foi realizado um teste de homogeneidade do qui-quadrado, que testa as hipóteses:

$$H_0\text{: Os diferentes } clusters \text{ têm distribuições homogéneas}$$

$$vs.$$

$$H_1\text{: Pelo menos um dos } clusters \text{ tem uma distribuição diferente dos restantes}$$

O objetivo será rejeitar a hipótese nula, o que nos garante que pelo um dos *clusters* será heterogéneo em relação aos restantes. Este teste funciona apenas como uma confirmação, não sendo estritamente necessário, uma vez que não nos garante que todos os *clusters* sejam heterogéneos entre si. No entanto, como o Método de Ward já realiza uma separação dos *clusters* por si só, não se sentiu a necessidade de aprofundar o teste à heterogeneidade dos diferentes grupos.

Como já foi referido, deu-se início às iterações do número de *clusters* com 5 grupos de setores. Para este valor, realizaram-se três testes individuais à homogeneidade, visto que dois dos *clusters* considerados incluíam apenas um setor (Extração e Atividades Financeiras, respetivamente), pelo facto de apresentarem taxas de incumprimento muito discrepantes das restantes, sendo os setores com maior e menor taxa de incumprimento, respetivamente. Note-se que estes setores se manterão, como é óbvio, sempre em *clusters* próprios ao longo das várias iterações. Assim, para os restantes três *clusters* criados, foi rejeitada a hipótese nula para um nível de significância de 5%, pelo que há evidência de que os setores incluídos nos três grupos não são homogéneos entre si. Note-se que, ao realizar o teste de homogeneidade geral, a hipótese nula foi rejeitada para todos os níveis de significância usuais, o que

reforça a ideia de falta de homogeneidade. Deste modo, procedeu-se à seguinte iteração, considerando um número de *clusters* igual a 6.

Considerando 6 grupos de setores, as conclusões foram as mesmas das obtidas para 5 *clusters*, já descritas acima, com a exceção de um *cluster* onde não se rejeitou a hipótese de homogeneidade. Porém, face aos restantes resultados, não se considerou isto suficiente e procedeu-se à iteração seguinte com a divisão em mais grupos.

Procedendo a uma divisão em 7 *clusters*, os resultados melhoraram, não se rejeitando a hipótese nula para um nível de significância de 5% em nenhum dos testes individuais, com a exceção de um grupo, que inclui os setores "*Holdings*", "Indústria Têxtil", "Transportes" "Comércio Grosso Alimentar" e "Agricultura". Quanto ao teste global de homogeneidade, também não se rejeitou a hipótese nula. Face a melhoria dos resultados, nesta iteração procedeu-se também ao teste da heterogeneidade entre grupos, rejeitando-se a hipótese nula, como se desejava.

Embora se tenha obtido uma melhoria nos resultados na iteração com 7 *clusters*, procedeu-se à iteração seguinte com 8 grupos, de modo a perceber se seria melhor optar por este número de *clusters*. Aconteceu o mesmo que na situação em que se considerou 7 *clusters*, ou seja, a hipótese nula do teste individual de homogeneidade foi rejeitada apenas para um grupo de setores, com a diferença de que, neste caso, o valor-P se situava nos 4%, valor muito próximo de 5%. O *cluster* em que isto aconteceu continha os setores "Indústria Têxtil", "Transportes", "Comércio Grosso Alimentar" e "Agricultura". Os restantes testes já mencionados apresentaram todos resultados positivos. Assim, como não é do melhor interesse ter demasiados *clusters*, visto que o objetivo é reduzir o número de categorias incluídas nesta variável, considerou-se os resultados obtidos para 8 grupos de setores, de um modo geral, bastante satisfatórios, pelo que se optou por esta iteração para criar a nova variável que traduz o setor de atividade da empresa.

Na Figura 4.3 apresenta-se, através do dendrograma, o agrupamento fornecido pelo Método de Ward pelo qual se optou, apresentando-se a constituição de cada *cluster* na Tabela 4.3.

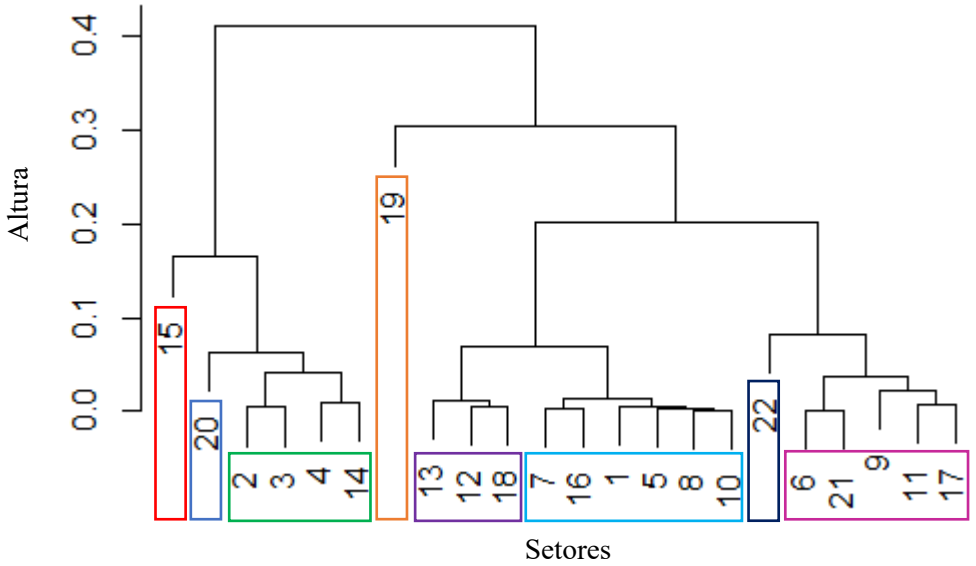


*Figura 4.3 – Dendrograma com a representação do agrupamento fornecido pelo Método de Ward selecionado*

*Tabela 4.3 - Constituição dos clusters obtidos pelo agrupamento selecionado após aplicação do Método de Ward.*

| Cluster | ID | Setor | Taxa de Incumprimento | Taxa de Incumprimento |
|---|---|---|---|---|
| **1** | 15 | Extração | 45,3% | 45,3% |
| **2** | 20 | *Holdings* | 36,6% | 36,6% |
| **3** | 2 | Indústria Têxtil | 33,2% | 31,7% |
| | 3 | Transportes | 32,8% | |
| | 4 | Comércio Grosso Alimentar | 29,6% | |
| | 14 | Agricultura | 30,6% | |
| **4** | 19 | Atividades Financeiras | 0,0% | 0,0% |
| **5** | 13 | Indústria Tecnológica | 26,9% | 27,6% |
| | 12 | Indústria Alimentar | 28,1% | |
| | 18 | Turismo de Restauração | 27,5% | |
| **6** | 7 | Indústria de Materiais | 24,4% | 23,8% |
| | 16 | *Utilities* | 24,7% | |
| | 1 | Indústria Transformadora | 24,0% | |
| | 5 | Serviços | 23,4% | |
| | 8 | Comércio Grosso de Materiais | 23,6% | |
| | 10 | Comércio de Automóveis | 23,7% | |
| **7** | 22 | Associações | 12,3% | 12,3% |
| **8** | 6 | Comércio de Retalho | 17,3% | 18,5% |
| | 21 | Estado | 17,3% | |
| | 9 | Indústria Química | 20,9% | |
| | 11 | Comércio Grosso de Equipamentos | 19,4% | |
| | 17 | Turismo de Hotelaria | 18,7% | |

Obteve-se assim uma nova variável com menos categorias para representar o setor de atividade de uma Média Empresa na presente amostra, que será utilizada aquando da construção dos modelos de previsão do incumprimento.

## 2. Poder de Negociação da Empresa

O poder de negociação da empresa pode ser caracterizado como alto, médio, baixo ou desconhecido.

*Tabela 4.4 - Distribuição da variável "Poder de Negociação da Empresa" pelas respetivas categorias*

| Poder de Negociação da Empresa | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Alto** | 4 881 | 1 030 | 5 911 | 17,4% |
| **Médio** | 12 932 | 4 004 | 16 936 | 23,6% |
| **Baixo** | 4 327 | 2 158 | 6 485 | 33,3% |
| **Desconhecido** | 360 | 308 | 668 | 46,1% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

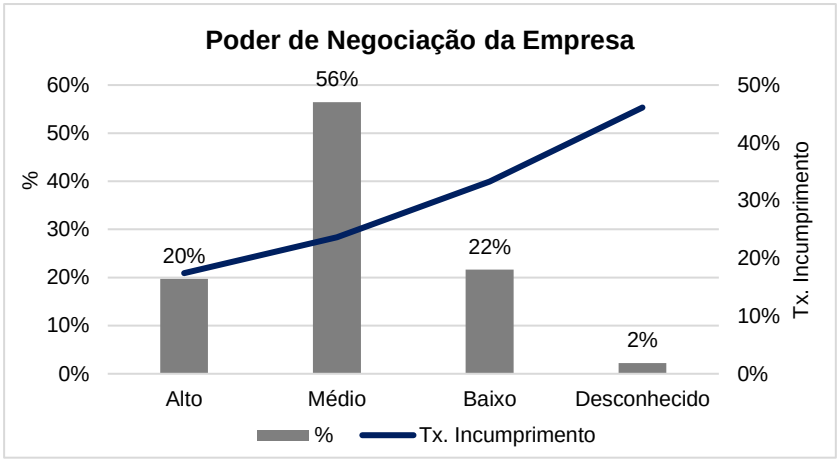


*Figura 4.4 – Poder de Negociação da Empresa: Distribuição e Taxas de Incumprimento*

A distribuição da taxa de incumprimento pelos vários tipos de poder de negociação é como seria de esperar: tão mais baixa quanto maior o poder de negociação da empresa. A taxa de incumprimento apresenta o maior valor quando esta informação é desconhecida, embora este caso represente apenas 2% das observações.

### 3. Periodicidade de Salários em Atraso

A variável que representa a periodicidade de salários em atraso de uma empresa responde à pergunta "A empresa já teve/tem salários dos seus empregados em atraso? Se sim, há quanto tempo?", podendo a resposta ser inserida numa das seguintes opções:

- Nunca;
- Nos últimos 3 meses;
- Há mais de 3 meses e até há 1 ano;
- Há mais de 1 ano e até há 3 anos;
- Há mais de 3 anos;
- Desconhecido.

*Tabela 4.5 - Distribuição da variável "Periodicidade de Salários em Atraso" pelas respetivas categorias*

| Periodicidade de Salários em Atraso | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Nunca** | 22 240 | 7 011 | 29 251 | 24,0% |
| **Nos últimos 3 meses** | 58 | 268 | 326 | 82,2% |
| **Há mais de 3 meses e até há 1 ano** | 9 | 15 | 24 | 62,5% |
| **Há mais de 1 ano e até há 3 anos** | 30 | 45 | 75 | 60,0% |
| **Há mais de 3 anos** | 4 | 25 | 29 | 86,2% |
| **Desconhecido** | 159 | 136 | 295 | 46,1% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

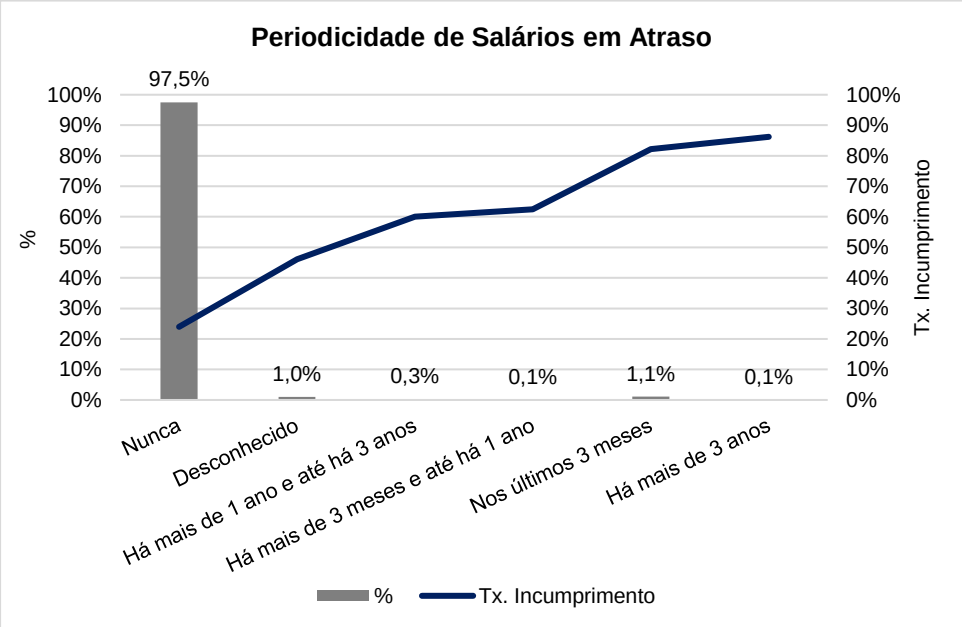


*Figura 4.5 – Periodicidade de Salários em Atraso: Distribuição e Taxas de Incumprimento*

Note-se que 97,5% das empresas da amostra nunca tiveram salários dos seus empregados em atraso, o que indica que esta variável pode não ser representativa. Porém, não se sabe se, nos casos em que teve, este facto foi determinante para o incumprimento futuro da empresa. De facto, o conjunto de empresas que nunca teve salários em atraso apresenta a menor taxa de incumprimento observada nesta variável, como seria de esperar. Quanto às restantes categorias, não se observa nenhum padrão de evolução da taxa de incumprimento à medida que o tempo há que se teve salários em atraso cresce, notando-se, mais uma vez, que estes casos têm uma representatividade de apenas 2,5% na amostra.

### 4. Dívidas em Atraso ao Estado

Em relação às dívidas em atraso da empresa ao Estado, esta pode não as ter, pode ter, dividindo-se esta opção no caso em que existe um plano de pagamento e no caso em que o mesmo não existe, ou esta informação pode ser desconhecida.

*Tabela 4.6 - Distribuição da variável "Dívidas em Atraso ao Estado" pelas respetivas categorias*

| Dívidas em Atraso ao Estado | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Não** | 21 340 | 5 960 | 27 300 | 21,8% |
| **Sim, com plano de pagamento** | 664 | 642 | 1 306 | 49,2% |
| **Sim, sem plano de pagamento** | 329 | 750 | 1 079 | 69,5% |
| **Desconhecido** | 167 | 148 | 315 | 47,0% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

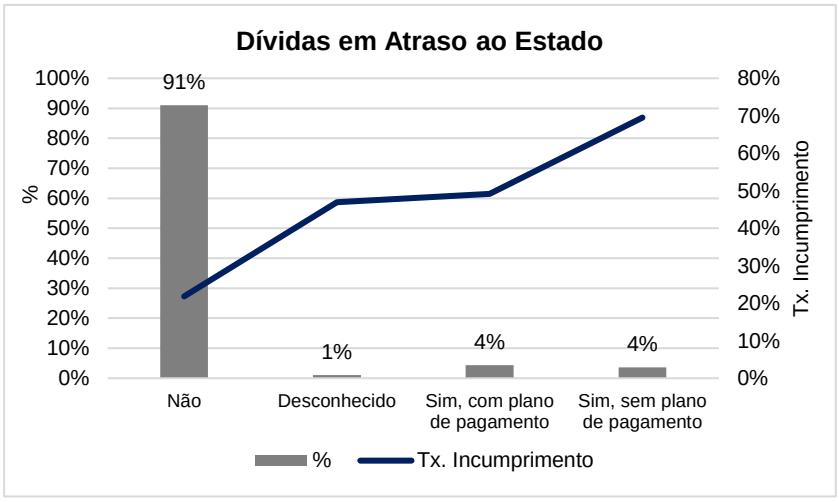


*Figura 4.6 – Dívidas em Atraso ao Estado: Distribuição e Taxas de Incumprimento*

Mais uma vez, a categoria da variável com menor taxa de incumprimento corresponde ao caso mais espectável, sendo, neste caso, o caso das Médias Empresas que não têm dívidas em atraso ao Estado. Observa-se ainda que, em caso afirmativo, a taxa de incumprimento é superior em casos em que não existe plano de pagamento, sendo também esta situação completamente expectável. Note-se que as empresas em que esta informação é desconhecida representam apenas 1% da amostra.

## 5. Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras

Em relação ao apoio dos sócios/acionistas principais da empresa em caso de necessidades financeiras, este pode ser elevado, médio, baixo, nulo ou desconhecido.

*Tabela 4.7 - Distribuição da variável "Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras" pelas respetivas categorias*

| Apoio Sócios/Acionistas Principais em Caso de Necessidades Financeiras | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Elevado** | 3 507 | 448 | 3 955 | 11,3% |
| **Médio** | 10 500 | 2 209 | 12 709 | 17,4% |
| **Baixo** | 4 082 | 2 230 | 6 312 | 35,3% |
| **Nulo** | 2 445 | 1 856 | 4 301 | 43,2% |
| **Desconhecido** | 1 966 | 757 | 2 723 | 27,8% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

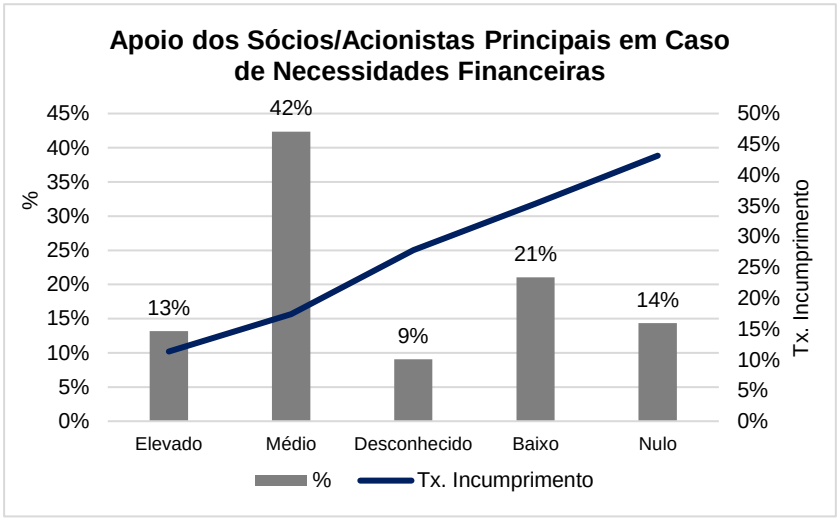


*Figura 4.7 – Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Distribuição e Taxas de Incumprimento*

Quanto menor o apoio dos sócios/acionistas principais, maior a taxa de incumprimento das empresas, como seria expectável. Observa-se ainda que 9% das observações da amostra correspondem a situações em que esta informação é desconhecida, obtendo-se para esta situação uma taxa de incumprimento de 27,8%, situando-se entre o nível de apoio médio e baixo.

## 6. Problemas entre Sócios

*Tabela 4.8 - Distribuição da variável "Problemas entre Sócios" pelas respetivas categorias*

| Problemas entre Sócios | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Desconhecido** | 171 | 148 | 319 | 46,4% |
| **Não** | 22 277 | 7 280 | 29 557 | 24,6% |
| **Sim** | 52 | 72 | 124 | 58,1% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

## 7. Problemas entre Sócios/Gerentes e Acionistas/Administradores

*Tabela 4.9 - Distribuição da variável "Problemas entre Sócios/Gerentes e Acionistas/Administradores" pelas respetivas categorias*

| Problemas entre Sócios/Gerentes e Acionistas/Administradores | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Desconhecido** | 155 | 136 | 291 | 46,7% |
| **Não** | 22 284 | 7 306 | 29 590 | 24,7% |
| **Sim** | 61 | 58 | 119 | 48,7% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

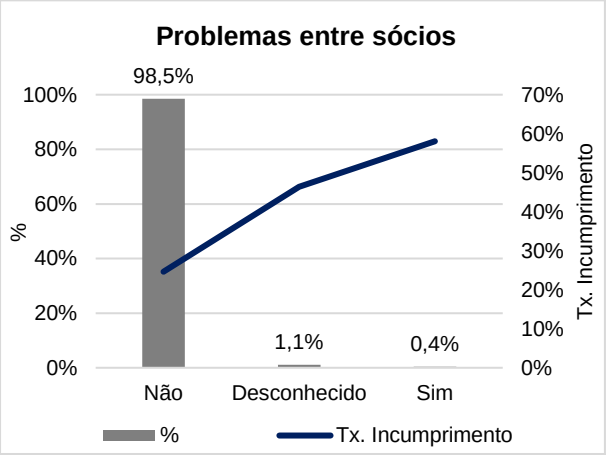


*Figura 4.8 – Problemas entre Sócios: Distribuição e Taxas de Incumprimento*

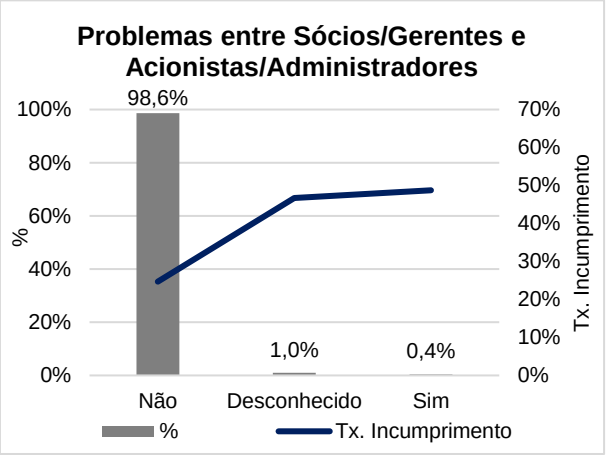


*Figura 4.9 – Problemas entre Sócios/Gerentes e Acionistas/Administradores: Distribuição e Taxas de Incumprimento*

Numa análise preliminar das variáveis, foi identificado que 99% das observações possuía o mesmo valor na variável que representa os problemas entre sócios e na variável que representa os problemas entre sócios/gerentes e acionistas/administradores. Depois de identificada esta situação, pretendeu-se selecionar apenas uma das variáveis para eventual entrada no modelo, uma vez que não faz sentido manter duas variáveis tão correlacionadas no mesmo modelo. Para escolher a variável a eliminar, construíram-se dois modelos de regressão logística simples com a variável resposta Def12m, tendo cada um dos modelos como variável independente cada uma das variáveis correlacionadas em análise. Depois de construídos os modelos, comparou-se os valores de $-2 \times ln\ (likelihood\ ratio)$ dos dois modelos, optando-se por manter a variável que representa os problemas entre sócios, por apresentar o maior valor.

*Tabela 4.10 – Seleção entre as variáveis "Problemas entre Sócios" e "Problemas entre Sócios/Gerentes e Acionistas/Administradores" através do likelihood ratio*

| | **Problemas entre Sócios** | **Problemas entre Sócios/Gerentes e Acionistas/Administradores** |
|---|---|---|
| $-2\ln\left(\frac{L_0}{L_1}\right)$ | 131,21 | 96,61 |

Note-se que a taxa de incumprimento mais baixa corresponde aos casos em que não existem problemas entre sócios e a mais alta aos casos em que existem estes problemas, como seria de esperar, encontrando-se num nível intermédio os casos em que esta informação é desconhecida, correspondendo, porém, a apenas 1% da amostra.

## 8. Capacidade de Substituição do Gestor Principal

Esta variável responde à questão "A empresa tem capacidade para substituir o seu gestor principal, em caso de necessidade?".

*Tabela 4.11 - Distribuição da variável "Capacidade de Substituição do Gestor Principal" pelas respetivas categorias*

| Capacidade de Substituição do Gestor Principal | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Desconhecido** | 221 | 212 | 433 | 49,0% |
| **Não** | 1 801 | 994 | 2 795 | 35,6% |
| **Sim** | 20 478 | 6 294 | 26 772 | 23,5% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

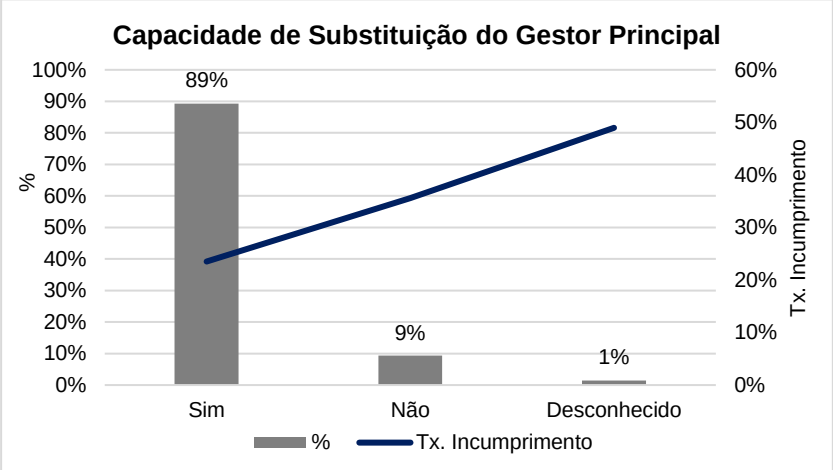


*Figura 4.10 – Capacidade de Substituição do Gestor Principal: Distribuição e Taxas de Incumprimento*

Tal como se tem verificado até ao momento com as restantes variáveis analisadas, o comportamento da taxa de incumprimento é o esperado, com as empresas sem capacidade de substituição com maior taxa de incumprimento do que as que têm essa capacidade. O caso em que esta informação é desconhecida apresenta a maior taxa de incumprimento, correspondendo a apenas 1% das observações da amostra.

## 9. Propriedade das Instalações

Esta variável caracteriza a relação da empresa com as suas instalações, conforme descrito na Tabela 4.12.

*Tabela 4.12 - Distribuição da variável "Propriedade das Instalações" pelas respetivas categorias*

| Propriedade das Instalações | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Próprias oneradas** | 3 305 | 1 600 | 4 905 | 32,6% |
| **Próprias não oneradas** | 10 227 | 2 239 | 12 466 | 18,0% |
| ***Leasing*** | 2 104 | 710 | 2 814 | 25,2% |
| **Arrendadas** | 4 869 | 2 036 | 6 905 | 29,5% |
| **Outros** | 1 995 | 915 | 2 910 | 31,4% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

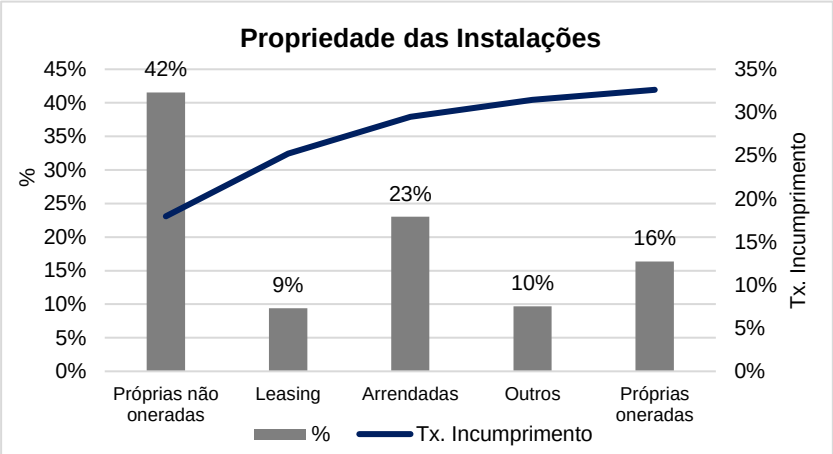


*Figura 4.11 – Propriedade das Instalações: Distribuição e Taxas de Incumprimento*

Observam-se maiores taxas de incumprimento nos casos em que as empresas possuem instalações próprias, mas oneradas, isto é, no caso em que existem encargos financeiros associados às mesmas, e nos casos classificados como "Outros", isto é, quando não se aplica nenhuma das outras categorias, apresentando taxas de incumprimento de cerca de 33% e 31%, respetivamente. Por outro lado, a menor taxa de incumprimento observa-se no caso de existência de instalações próprias não oneradas, com um valor de 18%, encontrando-se os casos de *leasing* e arrendamento em posições intermédias.

## 10. Redução/Renúncia de Linhas de Crédito

A redução ou renúncia de linhas de crédito de uma empresa num banco ocorre quando a instituição financeira decide diminuir ou encerrar uma ou mais linhas de crédito disponibilizadas à empresa. Essa decisão pode ser tomada por diversos motivos, como mudanças nas políticas internas do banco, redução do risco de crédito ou desempenho financeiro insatisfatório da empresa. Esta variável indica se esta situação ocorreu ou não para a empresa em questão.

*Tabela 4.13 - Distribuição da variável "Redução/Renúncia de Linhas de Crédito" pelas respetivas categorias*

| Redução/Renúncia de Linhas de Crédito | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Desconhecido** | 230 | 212 | 442 | 48,0% |
| **Não** | 22 128 | 7 002 | 29 130 | 24,0% |
| **Sim** | 142 | 286 | 428 | 66,8% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

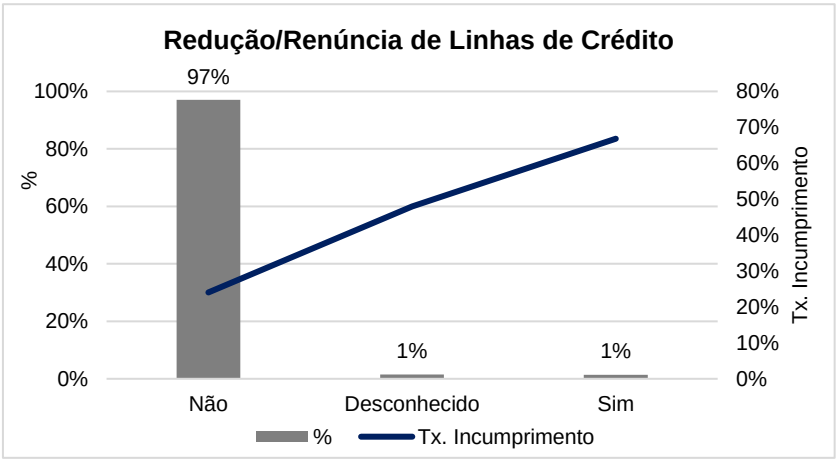


*Figura 4.12 – Redução/Renúncia de Linhas de Crédito: Distribuição e Taxas de Incumprimento*

Mais uma vez, o comportamento da variável em termos de taxa de incumprimento é o empiricamente esperado, existindo 1% da amostra em que esta informação é desconhecida.

### 11. Problemas de Pagamento aos Fornecedores/Credores

Em relação aos problemas de pagamento aos fornecedores/credores, a empresa pode não os ter, pode ter, dividindo-se esta opção no caso em que existe um plano de pagamento e no caso em que o mesmo não existe, ou esta informação pode ser desconhecida.

*Tabela 4.14 - Distribuição da variável "Problemas de Pagamento aos Fornecedores/Credores" pelas respetivas categorias*

| Problemas de Pagamento aos Fornecedores/Credores | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Sim, com plano de pagamento** | 928 | 1 323 | 2 251 | 58,8% |
| **Sim, sem plano de pagamento** | 349 | 734 | 1 083 | 67,8% |
| **Não** | 20 905 | 5 115 | 26 020 | 19,7% |
| **Desconhecido** | 318 | 328 | 646 | 50,8% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

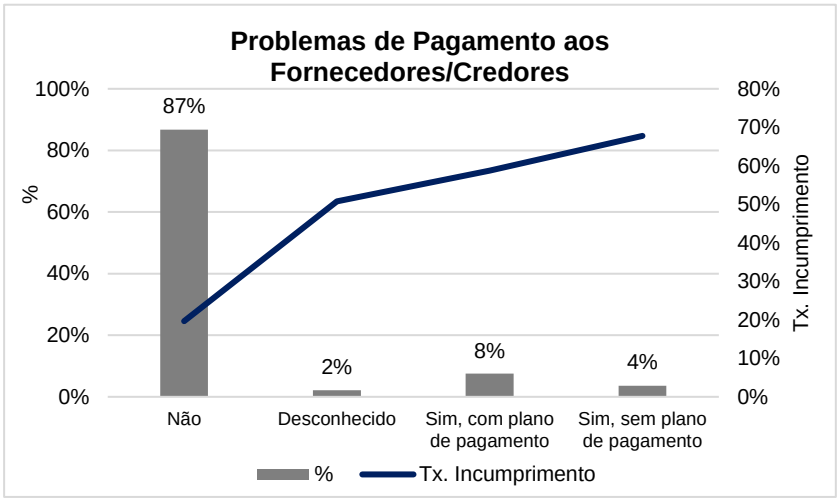


*Figura 4.13 – Problemas de Pagamento aos Fornecedores/Credores: Distribuição e Taxas de Incumprimento*

Como tem vindo a acontecer, a categoria da variável com menor taxa de incumprimento corresponde ao caso mais espectável, sendo, neste caso, o caso das Médias Empresas que não têm problemas de pagamento aos fornecedores/credores. Observa-se ainda que, em caso afirmativo, a taxa de incumprimento é superior em casos em que não existe plano de pagamento, sendo também esta situação completamente expectável. Note-se que as empresas em que esta informação é desconhecida representam apenas 2% da amostra.

## 12. Contas Auditadas no Ano de Observação

Esta variável fornece informação sobre se a empresa teve ou não contas auditadas no ano em observação e, em caso afirmativo, o tipo de auditoria efetuada.

*Tabela 4.15 - Distribuição da variável "Contas Auditadas no Ano de Observação" pelas respetivas categorias*

| Contas Auditadas no Ano de Observação | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Informação não disponível** | 2 647 | 1 342 | 3 989 | 33,6% |
| **Auditoria internacional** | 11 | 2 | 13 | 15,4% |
| **Auditoria nacional (sem reservas)** | 8 691 | 1 664 | 10 355 | 16,1% |
| **Auditoria nacional (com reservas)** | 1 983 | 1 632 | 3 615 | 45,1% |
| **Auditoria nacional (com reservas corrigidas)** | 256 | 98 | 354 | 27,7% |
| **Não auditada** | 8 912 | 2 762 | 11 674 | 23,7% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

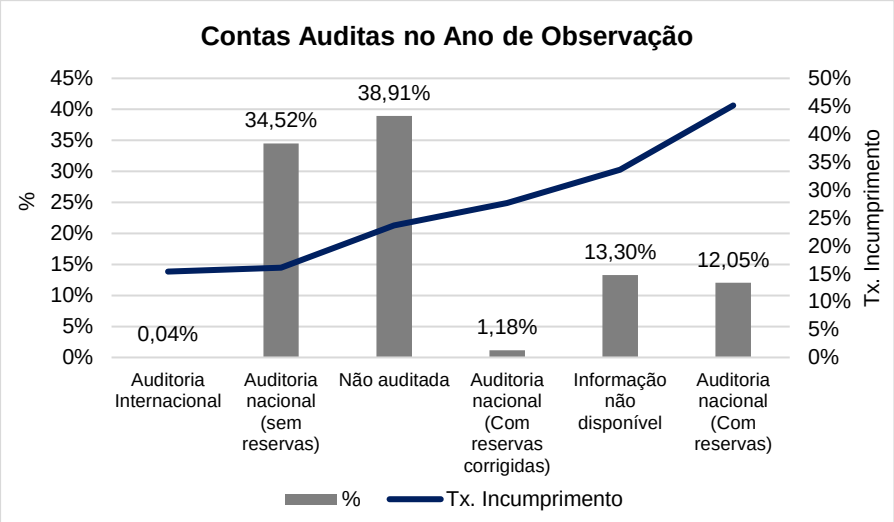


*Figura 4.14 – Contas Auditadas no Ano de Observação: Distribuição e Taxas de Incumprimento.*

Observa-se que os casos de auditorias internacionais não são representativos (0,04%). Quanto aos restantes casos, cerca de 13% das observações da amostra não possuem informação disponível, sendo a taxa de incumprimento destes casos relativamente elevada. Cerca de 48% das observações correspondem a empresas com auditorias nacionais no ano de observação, sendo a taxa de incumprimento superior quando há reservas e inferior quando estas não existem, como seria de esperar. Tem-se ainda que cerca de 39% das observações correspondem a empresas não auditadas no ano de observação, sendo a taxa de incumprimento deste caso de cerca de 24%, encontrando-se esta entre a dos casos de auditoria nacional sem reserva e a dos casos de auditoria nacional com reservas corrigidas.

### 13. Indicador de Rescisão de Cheque no Banco

Um indicador de rescisão de cheque é uma marcação feita pelo banco a um cheque que indica que o mesmo foi devolvido por algum motivo. Essa marcação pode indicar diversos motivos, como insuficiência de fundos na conta do emitente, cheque suspenso pelo próprio emitente, entre outros.

*Tabela 4.16 - Distribuição da variável "Indicador de Rescisão de Cheque no Banco" pelas respetivas categorias*

| Indicador de Rescisão de Cheque no Banco | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Desconhecido** | 4 675 | 1 085 | 5 760 | 18,8% |
| **Não** | 17 812 | 6 298 | 24 110 | 26,1% |
| **Sim** | 13 | 117 | 130 | 90,0% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

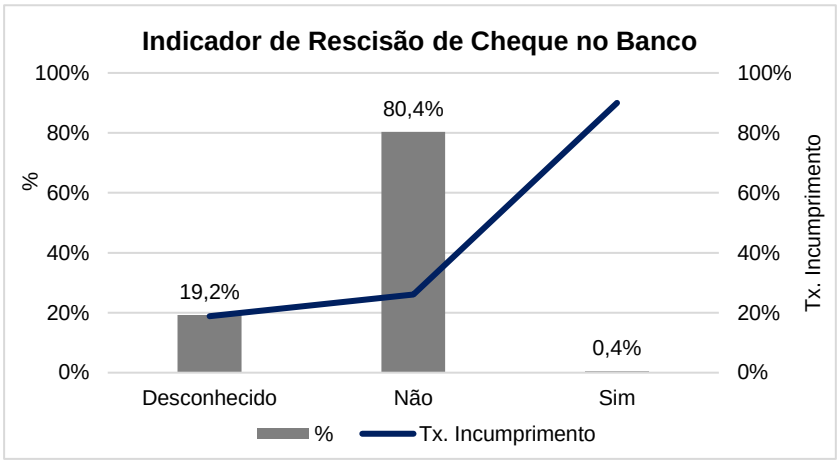


*Figura 4.15 – Indicador de Rescisão de Cheque no Banco: Distribuição e Taxas de Incumprimento*

Observa-se que 80% das observações da amostra correspondem ao caso em que não houve indicador de rescisão de cheque no banco, sendo a taxa de incumprimento inferior do que no caso afirmativo, como era expectável. Em caso afirmativo, a taxa de incumprimento é muito elevada, porém este valor está muito influenciado pelo facto de existirem poucas observações nesta situação (0,4%). Os restantes casos apresentam uma taxa de incumprimento mais baixa, correspondendo às situações em que esta informação é desconhecida.

### 14. Anos de Experiência do Gestor no Setor

Esta variável classifica num intervalo específico quantos anos de experiência tem o gestor da empresa no setor de atividade, conforme apresentado na Tabela 4.17.

*Tabela 4.17 - Distribuição da variável "Anos de Experiência do Gestor no Setor" pelas respetivas categorias*

| Anos de Experiência do Gestor no Setor | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **≤5** | 397 | 323 | 720 | 44,9% |
| **]5; 15]** | 2 335 | 1 071 | 3 406 | 31,4% |
| **]15; 25]** | 7 024 | 2 443 | 9 467 | 25,8% |
| **>25** | 12 744 | 3 663 | 16 407 | 22,3% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

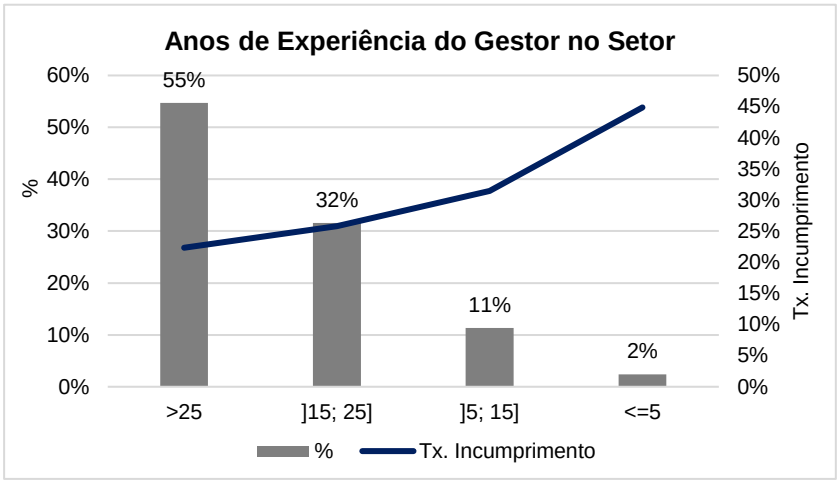


*Figura 4.16 – Anos de Experiência do Gestor no Setor: Distribuição e Taxas de Incumprimento*

O comportamento da variável é o expectável, isto é, a taxa de incumprimento é tão mais baixa quanto maior a experiência do gestor da média empresa no setor, sendo mais frequente existirem gestores com mais de 25 anos de experiência.

Inclui-se ainda neste subcapítulo, por ser também uma variável categórica, a variável que indica se a empresa se encontra ou não em moratória.

### 15. Indicador de Moratória

Esta variável indica, como já referido, se o contrato em questão se encontra em moratória no banco no momento da observação.

*Tabela 4.18 - Distribuição da variável "Indicador de Empresa em Moratória" pelas respetivas categorias*

| Indicador de Empresa em Moratória | Def12m=0 | Def12m=1 | Total | Taxa de Incumprimento |
|---|---|---|---|---|
| **Não** | 21011 | 6892 | 27 903 | 24,7% |
| **Sim** | 1489 | 608 | 2 097 | 29,0% |
| **Total** | **22 500** | **7 500** | **30 000** | **25%** |

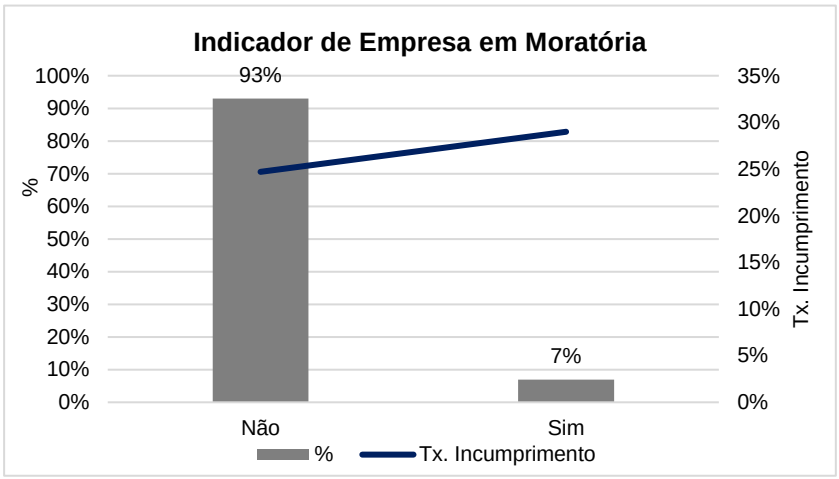


*Figura 4.17 – Indicador de Empresa em Moratória: Distribuição e Taxas de Incumprimento.*

À primeira vista, esta variável não parece ser muito importante, verificando-se que 93% das observações correspondem a casos em que a empresa não está em moratória e não se verificando uma grande discrepância entre as duas taxas de incumprimento apresentadas, embora a mesma seja superior em casos de moratória.

Esta análise das várias variáveis categóricas permitiu averiguar que a taxa de incumprimento nesta amostra apresenta o comportamento empiricamente esperado em relação às variáveis qualitativas disponíveis. Desta análise resultou um conjunto final de 14 variáveis categóricas potenciais a entrarem no modelo a ser construído.

#### 4.1.1.2. Variáveis Quantitativas

As variáveis quantitativas disponíveis na base de dados são as seguintes:

- Número de empregados;
- Antiguidade (em dias) da conta mais antiga ativa em que o sócio principal é o 1º titular;
- Montante total de crédito vencido no Banco;
- Total de crédito vencido há mais de 30 dias;
- Montante total de crédito vencido na Central de Riscos do Banco de Portugal;
- Montante total de crédito na Central de Riscos do Banco de Portugal;
- Resultado líquido do exercício no ano de observação;
- Resultados operacionais no ano de observação;
- Resultados operacionais no ano anterior;
- Amortizações acumuladas (amortizações no exercício) no ano de observação;
- Amortizações acumuladas (amortizações no exercício) no ano anterior;
- Provisões no ano de observação;
- Provisões no ano anterior;
- Custos com pessoal no ano de observação;
- Custos financeiros;
- Resultados financeiros;
- Total do ativo no ano de observação;
- Volume de faturação no ano de observação;
- Volume de faturação no ano anterior;
- Número de cheques devolvidos não justificados;
- *Cash-Flow*;
- Resultados correntes;
- Capitais Próprios;
- Total do passivo no ano de observação;
- Acréscimos e diferimentos (Ativo) no ano de observação;
- Acréscimos e diferimentos (Passivo) no ano de observação;
- Desconto comercial concedido (média dos últimos 12 meses);
- Saldo médio de devedores (média dos últimos 12 meses);
- Saldo médio de credores (média dos últimos 12 meses);
- Total de recursos (média dos últimos 12 meses);
- Total de responsabilidades (média dos últimos 12 meses).

Para averiguar se existem variáveis fortemente correlacionadas, foi calculado o coeficiente de correlação de Pearson para todas as variáveis quantitativas. Estabelecendo-se um *threshold* de 80%, identificou-se as variáveis apresentadas na Tabela 4.19 como fortemente correlacionadas:

*Tabela 4.19 – Variáveis quantitativas fortemente correlacionadas*

| Variável | Variável | Coeficiente de Correlação |
|---|---|---|
| Amortizações acumuladas (amortizações no exercício) no ano de observação | Amortizações acumuladas (amortizações no exercício) no ano anterior | 82% |
| Total do ativo no ano de observação | Total do passivo no ano de observação | 90% |
| Resultado líquido do exercício no ano de observação | Resultados operacionais no ano de observação | 91% |
| Volume de faturação no ano de observação | Volume de faturação no ano anterior | 95% |

Depois de identificadas estas correlações, é necessário escolher qual das duas variáveis altamente correlacionadas manter na base de dados, para eventualmente entrar no modelo a construir. Para tal, para cada par de variáveis foram criados dois modelos de regressão logística simples com a variável resposta do projeto, tendo cada um dos modelos como variável independente cada uma das variáveis correlacionadas em análise. Depois de construídos os modelos, comparou-se os valores dos $-2 \times ln\ (likelihood\ ratio)$ dos dois modelos, optando-se por manter a variável com o maior valor.

*Tabela 4.20 – Seleção entre as variáveis "Amortizações acumuladas (amortizações no exercício) no ano de observação" e "Amortizações acumuladas (amortizações no exercício) no ano anterior" através do likelihood ratio*

| | Amortizações acumuladas (amortizações no exercício) no ano de observação | Amortizações acumuladas (amortizações no exercício) no ano anterior |
|---|---|---|
| $-2\ln\left(\frac{L_0}{L_1}\right)$ | 9,26 | 6,97 |

*Tabela 4.21 – Seleção entre as variáveis "Total do ativo no ano de observação" e "Total do passivo no ano de observação" através do likelihood ratio*

| | Total do ativo no ano de observação | Total do passivo no ano de observação |
|---|---|---|
| $-2\ln\left(\frac{L_0}{L_1}\right)$ | 11,68 | 353,16 |

*Tabela 4.22 – Seleção entre as variáveis "Resultado líquido do exercício no ano de observação" e "Resultados operacionais no ano de observação" através do likelihood ratio*

| | Resultado líquido do exercício no ano de observação | Resultados operacionais no ano de observação |
|---|---|---|
| $-2\ln\left(\frac{L_0}{L_1}\right)$ | 1380,05 | 931,53 |

*Tabela 4.23 – Seleção entre as variáveis "Volume de faturação no ano de observação" e "Volume de faturação no ano anterior" através do likelihood ratio*

| | Volume de faturação no ano de observação | Volume de faturação no ano anterior |
|---|---|---|
| $-2\ln\left(\frac{L_0}{L_1}\right)$ | 83,54 | 45,74 |

Assim, optou-se por eliminar da base de dados, nesta análise preliminar das variáveis, as seguintes variáveis:

- Amortizações acumuladas (amortizações no exercício) no ano anterior;
- Total do ativo no ano de observação;
- Resultados operacionais no ano de observação;
- Volume de faturação no ano anterior.

Este tratamento resultou num conjunto final de 27 variáveis quantitativas.

### 4.1.2. Construção de Novas Variáveis: Rácios Financeiros e Comportamentais

Após análise e tratamento das variáveis originais da base de dados, optou-se por construir novas variáveis, com base nas variáveis quantitativas já existentes, construindo-se alguns rácios financeiros e comportamentais.

Os rácios financeiros construídos foram os seguintes, consoante as variáveis disponíveis na base de dados original para sua construção:

- $\frac{\text{Total do passivo no ano de observação}}{\text{Capitais Próprios}}$ (*Debt-to-Equity*);
- $\frac{\text{Resultado líquido do exercício no ano de observação}}{\text{Capitais Próprios}}$ (ROE);
- $\frac{\text{Capitais Próprios}}{\text{Total do ativo no ano de observação}}$ (Autonomia Financeira):
- $\frac{\text{Total do passivo no ano de observação} - \text{Acréscimos e diferimentos (Passivo) no ano de observação}}{\text{Total do ativo no ano de observação} - \text{Acréscimos e diferimentos (Ativo) no ano de observação}}$ (Alavancagem Financeira);
- $\frac{Cash-Flow}{\text{Custos financeiros}}$;
- $\frac{Cash-Flow}{\text{Total do ativo no ano de observação}}$.

Foram também construídos os seguintes três rácios comportamentais, que relacionam, respetivamente, os saldos de credores com os saldos de devedores, os recursos com as responsabilidades e as responsabilidades com o ativo:

- $\frac{\text{Saldo médio de credores (média dos últimos 12 meses)}}{\text{Saldo médio de devedores (média dos últimos 12 meses)}}$;
- $\frac{\text{Total de recursos (média dos últimos 12 meses)}}{\text{Total de responsabilidades (média dos últimos 12 meses)}}$;
- $\frac{\text{Total de responsabilidades (média dos últimos 12 meses)}}{\text{Total do ativo no ano de observação}}$.

Uma empresa não deve ter nem demasiadas dívidas a credores nem ter demasiadas dívidas a receber de devedores, esperando-se que haja um relativo equilíbrio entre ambos. Para além disso, espera-se que os recursos da empresa cubram as suas responsabilidades e, por outro lado, que o ativo cubra as responsabilidades.

A criação destes 9 rácios teve como objetivo construir novos indicadores que reflitam relações que possam eventualmente ser explicativas para o incumprimento futuro de uma empresa, através de informação que não esteja refletida na utilização imediata das variáveis financeiras originais.

Tal como foi efetuado para as variáveis quantitativas originais da base de dados, foram calculadas as correlações entre as várias variáveis, de modo a identificar se algumas se encontravam fortemente correlacionadas. Neste caso, para além de se efetuar as correlações apenas entre as novas variáveis construídas, incluiu-se também na matriz de correlação todas as variáveis originais utilizadas para a construção dos rácios, pois há uma possibilidade das mesmas terem correlação com variáveis a que deram origem.

Considerando o mesmo *threshold* já considerado anteriormente (80%) apenas se identificou um par de variáveis fortemente correlacionado: as variáveis que representam a autonomia financeira e a alavancagem financeira. Este par de variáveis apresenta um coeficiente de correlação de Pearson de -99,998%, sendo fortemente negativamente correlacionadas.

Como feito anteriormente com variáveis fortemente correlacionadas, foram construídos dois modelos de regressão logística simples.

*Tabela 4.24 – Seleção entre as variáveis "Autonomia Financeira" e "Alavancagem Financeira" através do likelihood ratio*

| | **Autonomia Financeira** | **Alavancagem Financeira** |
|---|---|---|
| $\mathbf{-2\ln\left(\frac{L_0}{L_1}\right)}$ | 27,33 | 28,26 |

A alavancagem apresenta maior valor de $-2 \times ln\ (likelihood\ ratio)$, logo optou-se por eliminar a variável que representa a autonomia financeira da empresa.

Assim, a criação destes rácios resultou num conjunto de mais oito variáveis quantitativas, terminando-se assim esta fase de análise e tratamento de dados com um conjunto de 14 variáveis categóricas/qualitativas e 35 variáveis quantitativas, sendo estas as variáveis candidatas a entrar nos modelos de regressão logística a criar nos capítulos que se seguem.

## 4.2.Construção e Avaliação de Modelos

### 4.2.1. Primeiro Modelo: Modelo Completo

Num momento inicial, optou-se por construir um modelo com todas as variáveis obtidas depois do respetivo tratamento realizado no capítulo 4.1, para averiguar a qualidade de um modelo que incluísse todas as 49 variáveis disponíveis. Para introdução no modelo, e em qualquer modelo construído neste projeto, todas as 14 variáveis categóricas foram transformadas em variáveis *dummy*. Esta transformação resultou em 46 variáveis *dummy*.

Obteve-se então, para o modelo completo, a curva ROC apresentada na Figura 4.18 e os resultados apresentados na Tabela 4.25.

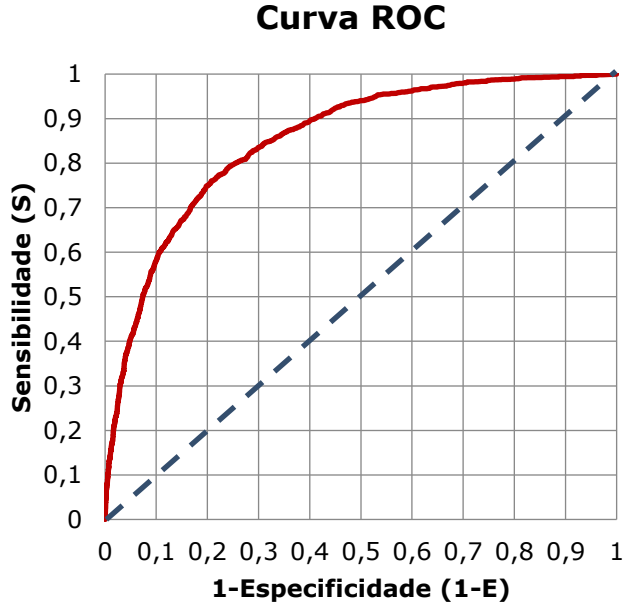


*Figura 4.18 - Curva ROC associada ao Modelo Completo*

*Tabela 4.25 - Resultados obtidos com o Modelo Completo*

| Fit Statistics | |
|---|---|
| -2 Log Likelihood | 24791 |
| AIC (smaller is better) | 24955 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 8949,5406 | 81 | <0,0001 |

Para além disto, obteve-se um valor de AUC (*area under the curve*) de 0,8525, sendo este valor bastante satisfatório.

Através do software SAS Enterprise Guide, foram obtidos ainda diversos pontos de corte e os respetivos valores necessários para a construção de uma matriz de confusão para cada ponto de corte. Como se pretende o maior equilíbrio possível entre a sensibilidade e a especificidade de um modelo, foi obtido o ponto correspondente à maior distância da curva ROC à bissetriz dos quadrantes ímpares apresentada na Figura 4.18, obtendo-se o ponto de corte de valor de 0,2555 para o modelo completo. Este ponto de corte indica-nos que, para este modelo, se uma observação obtiver uma probabilidade de incumprimento prevista inferior a este valor, então essa empresa é classificada como cumpridora, sendo classificada como incumpridora caso contrário. Utilizando este ponto de corte ótimo selecionado, construiu-se a matriz de confusão apresentada na Tabela 4.26 e respetivas sensibilidade e especificidade apresentadas na Tabela 4.27 (note-se que um valor positivo corresponde a uma empresa que entra em incumprimento).

*Tabela 4.26 - Matriz de Confusão associada ao Modelo Completo*

| | | **Observados** | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **5699** | 4669 | 10368 |
| | **Negativos** | 1801 | **17831** | 19632 |
| | **Total** | 7500 | 22500 | 30000 |

*Tabela 4.27 - Valores de Sensibilidade e Especificidade associados ao Modelo Completo*

| | |
|---|---|
| Sensibilidade | **75,99%** |
| Especificidade | **79,25%** |

Os valores de especificidade obtidos com este modelo são ambos bastante positivos, com especial destaque para a especificidade, que corresponde à percentagem de clientes que não entraram em incumprimento e foram corretamente classificados como cumpridores. Neste modelo, observou-se que 78,43% das observações obtiveram previsões corretas, tendo 21,57% sido mal classificadas, o que é um rácio bastante positivo.

Apesar do modelo completo apresentar resultados bastante positivos por si só, pretende-se obter um modelo parcimonioso, com o menor número de variáveis possíveis que expliquem o incumprimento, pelo que se realizou uma análise dos fatores de inflação da variância e se aplicou o método de seleção *stepwise*, conforme descrito no ponto seguinte.

### 4.2.2. Segundo Modelo: Aplicação de Fatores de Inflação da Variância e do Método de Seleção *Stepwise* ao Modelo Inicial

Após a construção do modelo inicial, pretendeu-se reduzir a dimensionalidade do conjunto das variáveis utilizadas no modelo.

Para apenas as variáveis quantitativas, a abordagem escolhida para tentar reduzir o respetivo número destas variáveis a entrar no modelo de regressão logística múltipla foi o da aplicação da análise dos fatores de inflação da variância (VIFs) ao conjunto das 35 variáveis quantitativas iniciais.

Nesta análise realizaram-se 9 iterações, tendo sido eliminadas 8 variáveis neste processo, considerando o critério de eliminação o de ter um valor de VIF correspondente superior a 3. As variáveis eliminadas foram as apresentadas abaixo, pela respetiva ordem, apresentando-se também o valor do respetivo fator de inflação da variância na iteração da sua eliminação:

- *Cash-Flow*: 12,67;
- Resultados correntes: 12,67;
- Resultado líquido do exercício no ano de observação: 8,83;
- Acréscimos e diferimentos (Ativo) no ano de observação: 5,19;
- Acréscimos e diferimentos (Passivo) no ano de observação: 5,19;
- Desconto comercial concedido (média dos últimos 12 meses): 4,21;
- Custos com pessoal no ano de observação: 3,51;
- Montante total de crédito vencido no Banco: 3,50.

Após esta eliminação de variáveis, construiu-se o modelo com as 14 variáveis categóricas iniciais (com todas as variáveis *dummy*) e com as restantes 27 variáveis contínuas, tendo sido aplicado o método de seleção *stepwise*, com o objetivo de redução de dimensionalidade já mencionado. Assim, em vez do total de 49 variáveis presentes no modelo completo, este modelo iniciou-se com 41 variáveis, tendo terminado com um número de variáveis ainda inferior devido ao método de seleção utilizado.

Após o método de seleção de variáveis, obteve-se um modelo que considera 20 das 27 variáveis quantitativas iniciais e 12 das 14 variáveis categóricas iniciais. Note-se, porém, que, considerando as variáveis *dummy,* que associam um coeficiente a cada um dos seus níveis, se passou de 82 (com 46 *dummies*) para 49 (com 28 *dummies*) coeficientes, incluindo o que corresponde ao termo constante.

Quanto às variáveis quantitativas, foram eliminadas as seguintes:

- Número de empregados;
- Resultados financeiros;
- $\frac{\text{Total do passivo no ano de observação}}{\text{Capitais Próprios}}$ (*Debt-to-Equity*);
- $\frac{\text{Resultado líquido do exercício no ano de observação}}{\text{Capitais Próprios}}$ (ROE);
- $\frac{Cash-Flow}{\text{Custos financeiros}}$;
- $\frac{\text{Saldo médio de credores (média dos últimos 12 meses)}}{\text{Saldo médio de devedores (média dos últimos 12 meses)}}$;
- $\frac{\text{Total de recursos (média dos últimos 12 meses)}}{\text{Total de responsabilidades (média dos últimos 12 meses)}}$.

Conclui-se, portanto, que a maior parte dos rácios construídos não são importantes para explicar o incumprimento das Médias Empresas, mantendo-se apenas os rácios

$$\frac{Cash-Flow}{\text{Total do ativo no ano de observação}}, \quad \frac{\text{Total de responsabilidades (média dos últimos 12 meses)}}{\text{Total do ativo no ano de observação}}$$

e

$$\frac{\text{Total do passivo no ano de observação} - \text{Acréscimos e diferimentos (Passivo) no ano de observação}}{\text{Total do ativo no ano de observação} - \text{Acréscimos e diferimentos (Ativo) no ano de observação}}$$

(Alavancagem Financeira).

Deste modo, quanto às variáveis quantitativas, permaneceram as seguintes no modelo:

- Antiguidade (em dias) da conta mais antiga ativa em que o sócio principal é o 1º titular;
- Total de crédito vencido há mais de 30 dias;
- Montante total de crédito vencido na Central de Riscos do Banco de Portugal;
- Montante total de crédito na Central de Riscos do Banco de Portugal;
- Resultados operacionais no ano anterior;
- Amortizações acumuladas (amortizações no exercício) no ano de observação;
- Provisões no ano de observação;
- Provisões no ano anterior;
- Custos financeiros;
- Volume de faturação no ano de observação;
- Número de cheques devolvidos não justificados;
- Capitais Próprios;

- Total do passivo no ano de observação;
- Saldo médio de devedores (média dos últimos 12 meses);
- Saldo médio de credores (média dos últimos 12 meses);
- Total de recursos (média dos últimos 12 meses);
- Total de responsabilidades (média dos últimos 12 meses);
- Alavancagem Financeira;
- $\frac{Cash-Flow}{\text{Ativo}}$;
- $\frac{\text{Responsabilidades}}{\text{Ativo}}$.

Quanto às variáveis categóricas, permaneceram no modelo as seguintes, com os níveis que também se indicam:

- Anos de Experiência do Gestor no Setor: >25;
- Anos de Experiência do Gestor no Setor: ]15; 25];
- Setor: *Cluster* 1 (Extração);
- Setor: *Cluster* 3 (Indústria Têxtil, Transportes, Comércio Grosso Alimentar, Agricultura);
- Setor: *Cluster* 5 (Indústria Tecnológica, Indústria Alimentar, Turismo de Restauração);
- Setor: *Cluster* 7 (Associações);
- Setor: *Cluster* 8 (Comércio de Retalho, Estado, Indústria Química, Comércio Grosso de Equipamentos, Turismo de Hotelaria);
- Poder de Negociação da Empresa: Alto;
- Poder de Negociação da Empresa: Médio;
- Periodicidade de Salários em Atraso: Nos últimos 3 meses;
- Periodicidade de Salários em Atraso: Há mais de 3 anos;
- Dívidas em Atraso ao Estado: Sim, com plano de pagamento;
- Dívidas em Atraso ao Estado: Sim, sem plano de pagamento;
- Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Elevado;
- Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Médio;
- Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Baixo;
- Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Nulo;
- Problemas entre Sócios: Sim;
- Capacidade de Substituição do Gestor Principal: Não;
- Propriedade das Instalações: Próprias não oneradas;
- Propriedade das Instalações: Arrendadas;
- Problemas de Pagamento aos Fornecedores/Credores: Não;
- Problemas de Pagamento aos Fornecedores/Credores: Desconhecido;
- Contas Auditadas no Ano de Observação: Informação não disponível;
- Contas Auditadas no Ano de Observação: Auditoria nacional (sem reservas);
- Contas Auditadas no Ano de Observação: Auditoria nacional (com reservas);
- Indicador de Rescisão de Cheque no Banco: Desconhecido;
- Indicador de Rescisão de Cheque no Banco: Não.

Obteve-se para este modelo a curva ROC e os resultados apresentados na Figura 4.19 e na Tabela 4.28.

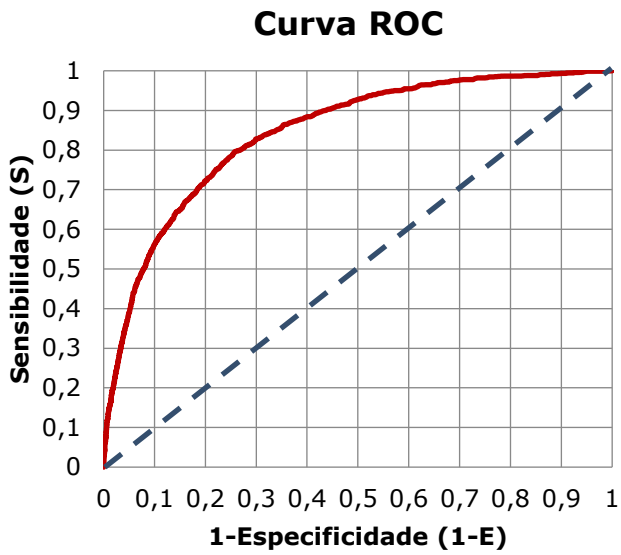


*Figura 4.19 - Curva ROC associada ao Segundo Modelo*

*Tabela 4.28 - Resultados obtidos com o Segundo Modelo*

| Fit Statistics | |
|---|---|
| -2 Log Likelihood | 25347 |
| AIC (smaller is better) | 25445 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 8393,6037 | 48 | <0,0001 |

Para além disto, obteve-se um valor de AUC de 0,8430, valor um pouco inferior ao obtido no modelo completo (0,8525). No entanto, uma *area under the curve* de 0,8430 é um resultado muito satisfatório.

Para este modelo, o ponto de corte ótimo selecionado foi de 0,2275, inferior ao obtido no modelo completo (0,2555). Com este ponto de corte, construiu-se a matriz de confusão e respetivas sensibilidade e especificidade, apresentadas na Tabela 4.29 e na Tabela 4.30, respetivamente.

*Tabela 4.29 - Matriz de Confusão associada ao Segundo Modelo*

| | | Observados | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **5964** | 5793 | 11757 |
| | **Negativos** | 1536 | **16707** | 18243 |
| | **Total** | 7500 | 22500 | 30000 |

*Tabela 4.30 - Valores de Sensibilidade e Especificidade associados ao Segundo Modelo*

| | |
|---|---|
| Sensibilidade | **79,52%** |
| Especificidade | **74,25%** |

Em comparação com o modelo completo, o valor da sensibilidade aumentou cerca de 3,5 pontos percentuais, sendo que a especificidade diminuiu cerca de 5 p.p.. Apesar do exposto, os valores continuam a ser muito satisfatórios, destacando-se agora a sensibilidade, que representa a percentagem de observações de empresas que entraram em incumprimento e foram corretamente classificadas como incumpridoras. Neste modelo, observou-se que 75,57% das observações obtiveram previsões corretas, tendo 24,43% sido mal classificadas, observando-se uma pequena degradação face ao primeiro modelo, que classificava corretamente 78,43% das observações.

Apesar de alguns indicadores piorarem em relação ao modelo inicial, o número de variáveis diminui consideravelmente com os métodos utilizados para a construção do segundo modelo, pelo que, uma vez que os decréscimos observados não são significativos, se considera este modelo melhor que o inicial. Destaca-se ainda que este modelo privilegia a correta classificação de incumpridores, o que é um ponto

positivo, já que, ainda que também seja importante classificar corretamente os cumpridores para que o banco não recuse a concessão de crédito que lhe possa ser benéfica, é talvez mais importante prever corretamente o incumprimento, de modo a prevenir perdas efetivas do banco.

Ainda que este modelo tenha menos variáveis, continua a ter 20 variáveis quantitativas, o que se considera um número bastante elevado. Com o objetivo de diminuir a dimensionalidade do conjunto de variáveis quantitativas, procedeu-se à realização de uma análise de componentes principais das mesmas, descrita no ponto seguinte, mantendo as variáveis categóricas já selecionadas.

### 4.2.3. Terceiro Modelo: Aplicação de Análise de Componentes Principais ao Segundo Modelo

Considerando-se as 20 variáveis quantitativas selecionadas no modelo anterior, procedeu-se a uma análise de componentes principais, com o auxílio do software SAS Enterprise Guide, obtendo-se os resultados apresentados na Tabela 4.31 e na Figura 4.20.

*Tabela 4.31 - Análise de Componentes Principais das variáveis quantitativas selecionadas no Segundo Modelo*

| Componente Principal | Valor Próprio | Proporção da Variância | Proporção da Variância Cumulativa |
|---|---|---|---|
| **1** | 3,57513279 | 0,1788 | 0,1788 |
| **2** | 1,86839729 | 0,0934 | 0,2722 |
| **3** | 1,48606942 | 0,0743 | 0,3465 |
| **4** | 1,28463418 | 0,0642 | 0,4107 |
| **5** | 1,24329273 | 0,0622 | 0,4729 |
| **6** | 1,01702207 | 0,0509 | 0,5237 |
| **7** | 1,0031962 | 0,0502 | 0,5739 |
| **8** | 0,99460357 | 0,0497 | 0,6236 |
| **9** | 0,96993553 | 0,0485 | 0,6721 |
| **10** | 0,95797495 | 0,0479 | 0,72 |
| **11** | 0,90919024 | 0,0455 | 0,7655 |
| **12** | 0,81126598 | 0,0406 | **0,806** |
| **13** | 0,76516507 | 0,0383 | 0,8443 |
| **14** | 0,6980503 | 0,0349 | 0,8792 |
| **15** | 0,59655485 | 0,0298 | 0,909 |
| **16** | 0,50002607 | 0,025 | 0,934 |
| **17** | 0,44581078 | 0,0223 | 0,9563 |
| **18** | 0,42422754 | 0,0212 | 0,9775 |
| **19** | 0,29000779 | 0,0145 | 0,992 |
| **20** | 0,15944266 | 0,008 | 1 |

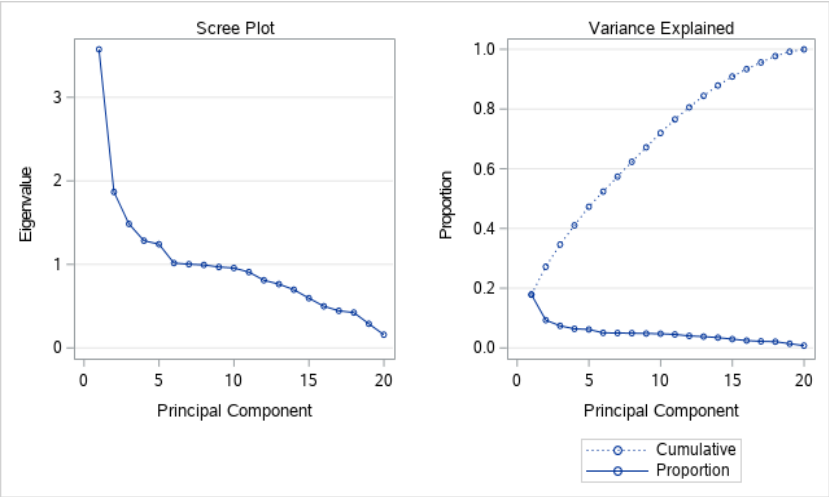


*Figura 4.20 – Valores Próprios e Proporção da Variância associados às Componentes Principais das variáveis quantitativas selecionadas no Segundo Modelo*

Para que o conjunto das componentes principais selecionadas para entrar no modelo expliquem pelo menos 80% da variância, foi necessário selecionar as primeiras 12 componentes principais. Com estas componentes principais e as variáveis categóricas já selecionadas pelo segundo modelo, criou-se um novo modelo de regressão logística múltipla, utilizando mais uma vez o método de seleção *stepwise*.

O método de seleção eliminou do modelo duas componentes principais, a 1ª e a 3ª, mantendo todas as variáveis categóricas já incluídas no segundo modelo (e as mesmas respetivas variáveis *dummy*). Assim, a aplicação da análise de componentes principais transformou o conjunto de 20 variáveis contínuas do segundo modelo em metade.

Obteve-se para este modelo a curva ROC e os resultados apresentados na Figura 4.21 e na Tabela 4.32, respetivamente.

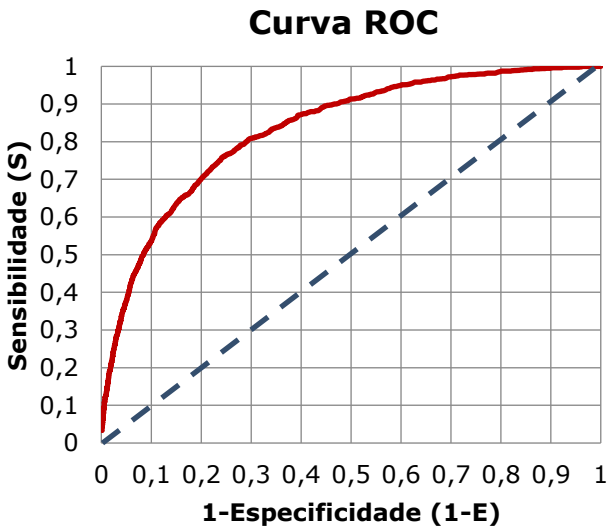


*Figura 4.21 - Curva ROC associada ao Terceiro Modelo*

*Tabela 4.32 - Resultados obtidos com o Terceiro Modelo*

| Fit Statistics | |
|---|---|
| -2 Log Likelihood | 25793 |
| AIC (smaller is better) | 25871 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 7946,8922 | 38 | <0,0001 |

Para além disto, obteve-se um valor de AUC de 0,8330, valor um pouco inferior ao obtido no segundo modelo (0,8430), embora, fora de uma perspetiva de comparação, bastante positivo. Esta diferença é pouco significativa.

Para este modelo, o ponto de corte ótimo selecionado foi de 0,235, sendo superior ao ponto de corte selecionado para o segundo (0,2275). Com este ponto de corte, construiu-se a matriz de confusão apresentada na Tabela 4.33 e respetivas sensibilidade e especificidade apresentadas na Tabela 4.34.

*Tabela 4.33 - Matriz de Confusão associada ao Terceiro Modelo*

| | | **Observados** | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **5710** | 5505 | 11215 |
| | **Negativos** | 1790 | **16995** | 18785 |
| | **Total** | 7500 | 22500 | 30000 |

*Tabela 4.34 - Valores de Sensibilidade e Especificidade associados ao Terceiro Modelo*

| | |
|---|---|
| Sensibilidade | **76,13%** |
| Especificidade | **75,53%** |

O valor da sensibilidade diminuiu cerca de 3,4 pontos percentuais em relação ao segundo modelo, sendo que a especificidade aumentou cerca de 1,3 p.p.. Apesar do exposto, os valores continuam a ser muito satisfatórios, existindo um excelente equilíbrio entre a sensibilidade e a especificidade, que, arredondando às unidades, apresentam ambos o valor de 76%. Neste modelo, observou-se que 75,68% das observações obtiveram previsões corretas, tendo 24,32% sido mal classificadas, observando-se um aumento de 0,11 p.p. na percentagem de classificações corretas em relação ao segundo modelo, verificando-se, embora mínima, uma melhoria.

Apesar de o valor de AUC ser inferior ao obtido para o segundo modelo, este terceiro modelo apresenta melhorias em termos de equilíbrio entre especificidade e sensibilidade e em termos de percentagem de previsões corretas. Para além disto, observa-se uma redução bastante significativa em termos de variáveis consideradas no modelo, passando-se de um conjunto de 32 variáveis no segundo modelo para um conjunto de 22 variáveis no terceiro, pelo que se considera este um melhor modelo do ponto de vista qualitativo.

## 4.3.Interpretação dos Modelos

Após a construção dos vários modelos, é importante interpretar os resultados obtidos, de modo a "traduzi-los" para o ponto de vista prático e de utilização no âmbito do negócio, isto é, é relevante entender quais os fatores que influenciam positivamente ou negativamente o incumprimento nos modelos obtidos. A interpretação do modelo completo não será apresentada, visto que, por não ser parcimonioso e existirem modelos de qualidade com menos variáveis como alternativa, não se considera este modelo relevante para as conclusões do projeto.

Para o modelo selecionado, isto é, o terceiro modelo, foram obtidos os resultados que se apresentam na Tabela 4.35.

*Tabela 4.35 - Terceiro Modelo: Estimador, Desvio-Padrão, Estatística de Teste e Valor-P associados a cada variável utilizada no modelo*

| Parâmetro | Estimador | Desvio-Padrão | Estatística de Teste | Valor-P |
|---|---|---|---|---|
| Termo Constante | 1,3113 | 0,3467 | 3,78 | 0,0002 |
| Anos Experiência do Gestor no Setor: >25 | -0,3876 | 0,05082 | -7,63 | <0,0001 |
| Anos Experiência do Gestor no Setor: ]5; 15] | -0,1861 | 0,04997 | -3,72 | 0,0002 |
| Setor: Cluster 1 | 0,9948 | 0,1665 | 5,98 | <0,0001 |
| Setor: Cluster 3 | 0,4431 | 0,03846 | 11,52 | <0,0001 |
| Setor: Cluster 5 | 0,2026 | 0,05666 | 3,58 | 0,0003 |
| Setor: Cluster 7 | -1,3011 | 0,2152 | -6,05 | <0,0001 |
| Setor: Cluster 8 | -0,4039 | 0,04406 | -9,17 | <0,0001 |
| Poder de Negociação da Empresa: Alto | -0,3555 | 0,05013 | -7,09 | <0,0001 |
| Poder de Negociação da Empresa: Médio | -0,1536 | 0,03747 | -4,1 | <0,0001 |
| Periodicidade de Salários em Atraso: Nos últimos 3 meses | 1,4862 | 0,1829 | 8,12 | <0,0001 |
| Periodicidade de Salários em Atraso: Há mais de 3 anos | 2,7411 | 0,6034 | 4,54 | <0,0001 |
| Dívidas em Atraso ao Estado: Sim, com plano de pagamento | 0,2366 | 0,07009 | 3,38 | 0,0007 |
| Dívidas em Atraso ao Estado: Sim, sem plano de pagamento | 0,7302 | 0,08412 | 8,68 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Elevado | -0,405 | 0,07611 | -5,32 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Médio | -0,2249 | 0,05939 | -3,79 | 0,0002 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Baixo | 0,3421 | 0,06128 | 5,58 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Nulo | 0,2607 | 0,06748 | 3,86 | 0,0001 |
| Problemas entre Sócios: Sim | 0,5706 | 0,2281 | 2,5 | 0,0124 |
| Capacidade de Substituição do Gestor Principal: Não | 0,4528 | 0,05009 | 9,04 | <0,0001 |
| Propriedade das Instalações: Próprias não oneradas | -0,2028 | 0,03776 | -5,37 | <0,0001 |
| Propriedade das Instalações: Arrendadas | 0,1993 | 0,04056 | 4,91 | <0,0001 |
| Problemas de Pagamento aos Fornecedores/Credores: Não | -1,1574 | 0,0484 | -23,91 | <0,0001 |
| Problemas de Pagamento aos Fornecedores/Credores: Desconhecido | -0,621 | 0,1109 | -5,6 | <0,0001 |
| Contas Auditadas no Ano de Observação: Informação não disponível | 0,3341 | 0,04658 | 7,17 | <0,0001 |
| Contas Auditadas no Ano de Observação: Auditoria nacional (sem reservas) | -0,2788 | 0,04046 | -6,89 | <0,0001 |
| Contas Auditadas no Ano de Observação: Auditoria nacional (com reservas) | 0,7801 | 0,04792 | 16,28 | <0,0001 |
| Indicador de Rescisão de Cheque no Banco: Desconhecido | -1,7413 | 0,3367 | -5,17 | <0,0001 |
| Indicador de Rescisão de Cheque no Banco: Sim | -1,2364 | 0,3352 | -3,69 | 0,0002 |
| Componente Principal 2 | -0,9552 | 0,03382 | -28,25 | <0,0001 |
| Componente Principal 4 | -0,2192 | 0,02431 | -9,02 | <0,0001 |
| Componente Principal 5 | -0,2769 | 0,02368 | -11,69 | <0,0001 |
| Componente Principal 6 | 1,0812 | 0,1067 | 10,13 | <0,0001 |
| Componente Principal 7 | -0,1126 | 0,02782 | -4,05 | <0,0001 |
| Componente Principal 8 | -0,394 | 0,04442 | -8,87 | <0,0001 |
| Componente Principal 9 | 0,678 | 0,06881 | 9,85 | <0,0001 |
| Componente Principal 10 | -0,4252 | 0,02907 | -14,63 | <0,0001 |
| Componente Principal 11 | 0,05955 | 0,02819 | 2,11 | 0,0347 |
| Componente Principal 12 | 0,2388 | 0,02687 | 8,89 | <0,0001 |

Analisando os resultados apresentados na Tabela 4.35, é possível interpretar quais as variáveis que influenciam de forma positiva ou negativa o incumprimento das empresas, individualmente, de acordo com o modelo selecionado (terceiro modelo).

Os fatores qualitativos que influenciam positivamente o incumprimento (com as respetivas variáveis destacadas a vermelho na Tabela 4.35), isto é, a sua ocorrência aumenta a probabilidade de a empresa em análise incumprir nos 12 meses seguintes, são os seguintes:

- O facto de a empresa pertencer aos setores: Extração, Indústria Têxtil, Transportes, Comércio Grosso Alimentar, Agricultura, Indústria Tecnológica, Indústria Alimentar ou Turismo de Restauração (consultar Tabela 4.3 para constituição de cada *cluster* da variável "Setor");
- O facto de a periodicidade de salários em atraso ser nos últimos 3 meses ou há mais de 3 anos;
- O facto de a empresa ter dívidas em atraso ao Estado, independentemente de ter ou não um plano de pagamento;
- O facto de o apoio dos sócios ou acionistas principais em caso de necessidades financeiras ser baixo ou nulo;
- A existência de problemas entre sócios;
- A incapacidade de substituição do gestor principal;
- O facto de as instalações da empresa serem arrendadas;
- A inexistência de informação relativa a contas auditadas no ano de observação;
- A existência de uma auditoria nacional com reservas no ano de observação.

Por outro lado, os fatores qualitativos que influenciam negativamente o incumprimento (com as respetivas variáveis destacadas a verde na Tabela 4.35), isto é, a sua ocorrência diminui a probabilidade da empresa em análise incumprir nos 12 meses seguintes, são os seguintes:

- O facto de o gestor da empresa ter entre 5 e 15 anos ou mais de 25 anos de experiência no setor;
- O facto de a empresa pertencer aos setores: Associações, Comércio de Retalho, Estado, Indústria Química, Comércio Grosso de Equipamentos ou Turismo de Hotelaria (consultar Tabela 4.3 para constituição de cada cluster da variável "Setor");
- O facto de a empresa ter um poder de negociação alto ou médio;
- O facto de o apoio dos sócios ou acionistas principais em caso de necessidades financeiras ser elevado ou médio;
- O facto de as instalações da empresa serem próprias não oneradas;
- O facto de não existirem ou de não se possuir informação relativamente a problemas de pagamento aos fornecedores ou credores;
- A existência de uma auditoria nacional sem reservas no ano de observação;
- A existência ou falta de informação relativamente a um indicador de rescisão de cheque no banco.

Quanto aos fatores qualitativos, é possível observar que os mesmos influenciam o incumprimento de uma empresa da forma empiricamente esperada, com a exceção da inexistência de informação relativamente a problemas de pagamento aos fornecedores ou credores e da existência ou falta de informação relativamente a um indicador de rescisão de cheque no banco.

Quanto aos fatores quantitativos, apenas se possui informação relativamente a como cada componente principal criada que entrou no modelo influencia o incumprimento. Assim, as componentes principais 6, 9, 11 e 12 influenciam positivamente o incumprimento, isto é, quanto maior for o seu valor, maior a probabilidade de a empresa em análise incumprir nos 12 meses seguintes. Por outro lado, as

componentes principais 2, 4, 5, 7, 8 e 10 influenciam negativamente o incumprimento, isto é, quanto maior for o seu valor, menor a probabilidade da empresa em análise incumprir nos 12 meses seguintes.

Quanto a estas componentes principais, a Tabela 4.36 mostra como se correlacionam com as variáveis originais.

*Tabela 4.36 - Correlação entre as Componentes Principais e as respetivas variáveis originais*

| | CP2 | CP4 | CP5 | CP6 | CP7 | CP8 | CP9 | CP10 | CP11 | CP12 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Antiguidade (em dias) da conta mais antiga ativa em que o sócio principal é o 1º titular** | 0,09 | 0,14 | -0,05 | 0,36 | 0,21 | -0,17 | -0,40 | 0,67 | -0,09 | -0,10 |
| **Total de crédito vencido há mais de 30 dias** | -0,45 | -0,05 | 0,01 | -0,15 | 0,00 | 0,04 | -0,04 | -0,02 | 0,07 | 0,06 |
| **Montante total de crédito vencido na Central de Riscos do Banco de Portugal** | -0,45 | -0,02 | 0,01 | 0,03 | -0,02 | -0,01 | -0,03 | -0,03 | 0,00 | 0,10 |
| **Montante total de crédito na Central de Riscos do Banco de Portugal** | -0,12 | -0,13 | -0,09 | 0,01 | 0,06 | -0,14 | -0,04 | 0,19 | 0,24 | 0,26 |
| **Resultados operacionais no ano anterior** | 0,41 | -0,51 | -0,33 | -0,03 | 0,00 | 0,13 | 0,16 | 0,11 | -0,02 | 0,03 |
| **Amortizações acumuladas (amortizações no exercício) no ano de observação** | 0,11 | -0,09 | 0,02 | 0,04 | -0,03 | 0,01 | -0,07 | -0,16 | -0,18 | -0,30 |
| **Provisões no ano de observação** | -0,14 | 0,15 | 0,21 | 0,00 | -0,04 | 0,55 | 0,51 | 0,45 | -0,32 | 0,02 |
| **Provisões no ano anterior** | -0,20 | 0,59 | 0,47 | 0,02 | 0,00 | -0,10 | -0,09 | -0,09 | -0,05 | -0,05 |
| **Custos financeiros** | -0,34 | -0,02 | 0,05 | -0,10 | -0,02 | 0,01 | 0,05 | -0,14 | -0,09 | -0,06 |
| **Volume de faturação no ano de observação** | 0,21 | -0,13 | 0,00 | 0,09 | -0,02 | 0,00 | -0,08 | -0,04 | -0,12 | -0,36 |
| **Número de cheques devolvidos não justificados** | -0,08 | 0,01 | -0,04 | 0,78 | -0,10 | -0,28 | 0,49 | -0,17 | 0,07 | 0,01 |
| **Capitais Próprios** | 0,43 | 0,12 | -0,05 | 0,08 | -0,07 | -0,05 | -0,16 | -0,19 | -0,57 | 0,57 |
| **Total do passivo no ano de observação** | -0,16 | -0,02 | 0,02 | -0,05 | -0,01 | -0,04 | 0,04 | -0,13 | -0,10 | -0,04 |
| **Saldo médio de devedores (média dos últimos 12 meses)** | -0,08 | -0,02 | 0,00 | 0,43 | -0,13 | 0,69 | -0,40 | -0,19 | 0,27 | 0,11 |
| **Saldo médio de credores (média dos últimos 12 meses)** | 0,61 | 0,33 | 0,03 | -0,07 | 0,01 | 0,05 | 0,11 | 0,03 | 0,32 | -0,05 |
| **Total de recursos (média dos últimos 12 meses)** | 0,57 | 0,40 | 0,06 | -0,09 | -0,01 | 0,05 | 0,12 | -0,08 | 0,21 | 0,03 |
| **Total de responsabilidades (média dos últimos 12 meses)** | -0,21 | 0,04 | -0,03 | -0,09 | 0,06 | -0,10 | 0,11 | 0,16 | 0,31 | 0,38 |
| **Alavancagem Financeira** | -0,01 | 0,01 | -0,02 | 0,07 | 0,95 | 0,15 | 0,09 | -0,22 | -0,04 | 0,02 |
| **Cash-Flow/Ativo** | 0,26 | -0,40 | 0,63 | -0,04 | 0,00 | 0,00 | 0,07 | 0,02 | 0,17 | 0,09 |
| **Responsabilidades/Ativo** | -0,13 | 0,38 | -0,67 | -0,11 | -0,06 | 0,10 | 0,11 | -0,03 | 0,04 | -0,06 |

Destaca-se na Tabela 4.36, para cada componente principal (CP), as duas variáveis originais com maior correlação positiva com a mesma (a verde) e as duas variáveis originais com maior correlação negativa com a mesma (a vermelho).

Apesar da qualidade deste modelo e do facto de ser parcimonioso e de mais fácil utilização, existe uma limitação quanto à interpretabilidade das variáveis contínuas que foram substituídas pelas respetivas componentes principais. Embora se saiba como cada componente principal influencia a probabilidade de incumprimento de uma empresa e como cada componente principal se relaciona com as variáveis originais, torna-se complicado estabelecer uma relação direta entre a variável original e a probabilidade de incumprimento, uma vez que cada componente principal resulta de uma combinação linear de várias variáveis, traduzindo assim a interação entre as várias variáveis e não apenas o comportamento de uma variável individualmente.

Assim, considera-se relevante apresentar também a interpretação do segundo modelo, que em tudo é semelhante ao modelo acima descrito, com a exceção de as variáveis contínuas não terem sido substituídas pelas respetivas componentes principais. A Tabela 4.37 apresenta os resultados do ajustamento deste modelo.

*Tabela 4.37 - Segundo Modelo: Estimador, Desvio-Padrão, Estatística de Teste e Valor-P associados a cada variável utilizada no modelo*

| Parâmetro | Estimador | Desvio-Padrão | Estatística de Teste | Valor-P |
|---|---|---|---|---|
| Termo Constante | 1,8374 | 0,3456 | 5,32 | <0,0001 |
| Anos Experiência do Gestor no Setor: >25 | -0,4061 | 0,05128 | -7,92 | <0,0001 |
| Anos Experiência do Gestor no Setor: ]5; 15] | -0,1859 | 0,05032 | -3,7 | 0,0002 |
| Setor: Cluster 1 | 1,332 | 0,173 | 7,7 | <0,0001 |
| Setor: Cluster 3 | 0,4618 | 0,03899 | 11,84 | <0,0001 |
| Setor: Cluster 5 | 0,2399 | 0,05738 | 4,18 | <0,0001 |
| Setor: Cluster 7 | -1,1718 | 0,2092 | -5,6 | <0,0001 |
| Setor: Cluster 8 | -0,3897 | 0,04423 | -8,81 | <0,0001 |
| Poder de Negociação da Empresa: Alto | -0,3431 | 0,05057 | -6,79 | <0,0001 |
| Poder de Negociação da Empresa: Médio | -0,155 | 0,0378 | -4,1 | <0,0001 |
| Periodicidade de Salários em Atraso: Nos últimos 3 meses | 1,3514 | 0,1759 | 7,68 | <0,0001 |
| Periodicidade de Salários em Atraso: Há mais de 3 anos | 2,6171 | 0,5945 | 4,4 | <0,0001 |
| Dívidas em Atraso ao Estado: Sim, com plano de pagamento | 0,2746 | 0,07058 | 3,89 | 0,0001 |
| Dívidas em Atraso ao Estado: Sim, sem plano de pagamento | 0,725 | 0,08429 | 8,6 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Elevado | -0,4163 | 0,0768 | -5,42 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Médio | -0,2295 | 0,05984 | -3,84 | 0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Baixo | 0,3167 | 0,06175 | 5,13 | <0,0001 |
| Apoio dos Sócios/Acionistas Principais em Caso de Necessidades Financeiras: Nulo | 0,2575 | 0,06813 | 3,78 | 0,0002 |
| Problemas entre Sócios: Sim | 0,6114 | 0,2259 | 2,71 | 0,0068 |
| Capacidade de Substituição do Gestor Principal: Não | 0,4301 | 0,05042 | 8,53 | <0,0001 |
| Propriedade das Instalações: Próprias não oneradas | -0,2202 | 0,03831 | -5,75 | <0,0001 |
| Propriedade das Instalações: Arrendadas | 0,1771 | 0,04111 | 4,31 | <0,0001 |
| Problemas de Pagamento aos Fornecedores/Credores: Não | -1,1228 | 0,04926 | -22,79 | <0,0001 |
| Problemas de Pagamento aos Fornecedores/Credores: Desconhecido | -0,5305 | 0,1133 | -4,68 | <0,0001 |
| Contas Auditadas no Ano de Observação: Informação não disponível | 0,3407 | 0,04668 | 7,3 | <0,0001 |
| Contas Auditadas no Ano de Observação: Auditoria nacional (sem reservas) | -0,2117 | 0,04359 | -4,86 | <0,0001 |
| Contas Auditadas no Ano de Observação: Auditoria nacional (com reservas) | 0,7761 | 0,05033 | 15,42 | <0,0001 |
| Indicador de Rescisão de Cheque no Banco: Desconhecido | -1,6858 | 0,3362 | -5,01 | <0,0001 |
| Indicador de Rescisão de Cheque no Banco: Sim | -1,2064 | 0,3347 | -3,6 | 0,0003 |
| Antiguidade (em dias) da conta mais antiga ativa em que o sócio principal é o 1º titular | -0,00007 | 5,97E-06 | -11,57 | <0,0001 |
| Total de crédito vencido há mais de 30 dias | -2,58E-06 | 0 | ∞ | <0,0001 |
| Montante total de crédito vencido na Central de Riscos do Banco de Portugal | 2,25E-06 | 0 | ∞ | <0,0001 |
| Montante total de crédito na Central de Riscos do Banco de Portugal | 9,44E-08 | 0 | ∞ | <0,0001 |
| Resultados operacionais no ano anterior | -3,11E-07 | 0 | ∞ | <0,0001 |
| Amortizações acumuladas (amortizações no exercício) no ano de observação | -7,65E-07 | 0 | ∞ | <0,0001 |
| Provisões no ano de observação | -1,34E-07 | 0 | ∞ | <0,0001 |

| Parâmetro | Estimador | Desvio-Padrão | Estatística de Teste | Valor-P |
|---|---|---|---|---|
| Provisões no ano anterior | -5,41E-07 | 0 | ∞ | <0,0001 |
| Custos financeiros | 1,29E-06 | 0 | ∞ | <0,0001 |
| Volume de faturação no ano de observação | -1,56E-08 | 0 | ∞ | <0,0001 |
| Número de cheques devolvidos não justificados | 0,5708 | 0,05923 | 9,64 | <0,0001 |
| Capitais Próprios | -3,89E-08 | 0 | ∞ | <0,0001 |
| Total do passivo no ano de observação | 2,01E-08 | 0 | ∞ | <0,0001 |
| Saldo médio de devedores (média dos últimos 12 meses) | 2,03E-06 | 0 | ∞ | <0,0001 |
| Saldo médio de credores (média dos últimos 12 meses) | -4,42E-06 | 0 | ∞ | <0,0001 |
| Total de recursos (média dos últimos 12 meses) | -2,42E-07 | 0 | ∞ | <0,0001 |
| Total de responsabilidades (média dos últimos 12 meses) | 5,06E-08 | 0 | ∞ | <0,0001 |
| Alavancagem Financeira | 0,0004 | 0,000114 | 3,5 | 0,0005 |
| Cash-Flow/Ativo | -1,2214 | 0,1311 | -9,32 | <0,0001 |
| Responsabilidades/Ativo | 0,03894 | 0,01868 | 2,08 | 0,0372 |

Em termos de fatores qualitativos, verifica-se que as variáveis apresentam o mesmo comportamento já observado para o terceiro modelo.

Quanto aos fatores quantitativos, os que influenciam positivamente o incumprimento de acordo com o segundo modelo (com as respetivas variáveis destacadas a vermelho na Tabela 4.37), isto é, quanto maior for o seu valor, maior a probabilidade de a empresa em análise incumprir nos 12 meses seguintes, são os seguintes:

- Montante total de crédito vencido na Central de Riscos do Banco de Portugal;
- Montante total de crédito na Central de Riscos do Banco de Portugal;
- Custos financeiros;
- Número de cheques devolvidos não justificados;
- Total do passivo no ano de observação;
- Saldo médio de devedores (média dos últimos 12 meses);
- Total de responsabilidades (média dos últimos 12 meses)
- Rácio de Alavancagem Financeira;
- Rácio $\frac{\text{Responsabilidades}}{\text{Ativo}}$.

Por outro lado, os fatores quantitativos que influenciam negativamente o incumprimento de acordo com o segundo modelo (com as respetivas variáveis destacadas a verde na Tabela 4.37), isto é, quanto maior for o seu valor, menor a probabilidade de a empresa em análise incumprir nos 12 meses seguintes, são os seguintes:

- Antiguidade (em dias) da conta mais antiga ativa em que o sócio principal é o 1º titular;
- Total de crédito vencido há mais de 30 dias;
- Resultados operacionais no ano anterior;
- Amortizações acumuladas (amortizações no exercício) no ano de observação;
- Provisões no ano de observação;
- Provisões no ano anterior;
- Volume de faturação no ano de observação;
- Capitais Próprios;
- Saldo médio de credores (média dos últimos 12 meses);

- Total de recursos (média dos últimos 12 meses);
- Rácio $\frac{\text{Cash-Flow}}{\text{Ativo}}$.

Na sua maioria, o comportamento das variáveis quantitativas no segundo modelo é o empiricamente esperado.

Deste modo, embora neste projeto se tenha optado por selecionar o terceiro e último modelo criado, pelo facto de o objetivo ter sido sempre obter um modelo parcimonioso, de modo a obter o mínimo erro de estimação possível e para que a aplicação do modelo seja mais simples, apresenta-se como alternativa o segundo modelo, para o caso em que o utilizador dos modelos prefira a utilização direta das variáveis absolutamente contínuas. Note-se, porém, que esta eventual opção pelo modelo sem a utilização da análise de componentes principais resulta na utilização de um modelo com mais 10 variáveis do que na alternativa já mencionada.

Em jeito de resumo, a Tabela 4.38 apresenta uma comparação entre os dois melhores modelos construídos:

*Tabela 4.38 - Comparação entre o segundo e o terceiro modelo*

| Indicadores | Segundo Modelo (sem CP) | Terceiro Modelo (com CP) | Modelo Mais Vantajoso |
|---|---|---|---|
| **Ponto de Corte** | 0,2275 | 0,235 | As diferenças entre os modelos não são significativas. |
| **AUC** | 0,8430 | 0,8330 | |
| **Sensibilidade** | 79,52% | 76,13% | |
| **Especificidade** | 74,25% | 75,53% | |
| **Percentagem de Previsões Corretas** | 77,57% | 75,68% | |
| **Número de Variáveis** | 32 | 22 | Modelo 3 |
| **Interpretabilidade** | Todas as variáveis podem ser diretamente interpretadas. | As variáveis correspondentes a componentes principais não são diretamente interpretáveis. | Modelo 2 |

## 4.4.Aplicação do Melhor Modelo

Com o objetivo de testar a qualidade do modelo construído e selecionado neste projeto, foi recolhida uma amostra independente da amostra utilizada para construção dos modelos, para aplicar o modelo final à mesma e avaliar a qualidade dos resultados obtidos. Foi recolhida uma amostra com 14.999 observações, não existindo qualquer interseção entre esta nova amostra e a base de dados utilizada ao longo do projeto.

Procedeu-se à aplicação do modelo final a esta amostra, obtendo-se a curva ROC ilustrada na Figura 4.22.

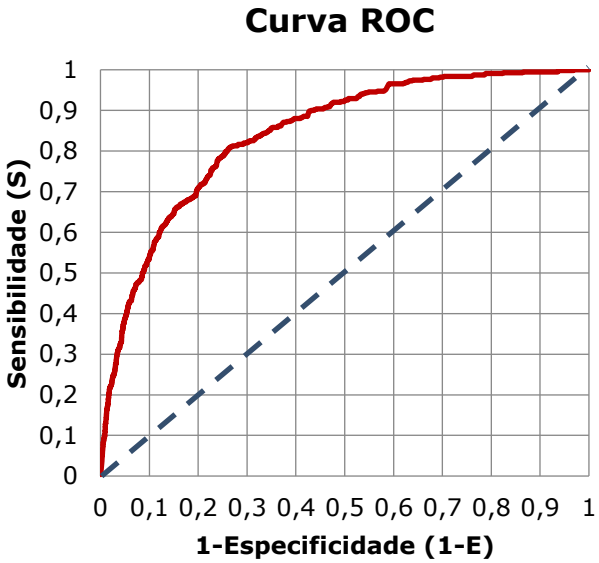


*Figura 4.22 - Curva ROC obtida após aplicação do modelo selecionado a uma amostra independente*

Obteve-se uma área abaixo da curva no valor de 0,8413, sendo este valor superior à AUC obtida aquando da construção do próprio modelo. Utilizando o ponto de corte selecionado para este modelo, que era de 0,235, construiu-se a respetiva matriz de confusão e obteve-se os valores de sensibilidade e especificidade para esta amostra.

*Tabela 4.39 - Matriz de Confusão obtida após aplicação do modelo selecionado a uma amostra independente*

| | | **Observados** | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **451** | 4258 | 4709 |
| | **Negativos** | 99 | **10191** | 10290 |
| | **Total** | 550 | 14449 | 14999 |

*Tabela 4.40 - Valores de Sensibilidade e Especificidade obtidos após aplicação do modelo selecionado a uma amostra independente*

| | |
|---|---|
| Sensibilidade | **82,00%** |
| Especificidade | **70,53%** |

Como se pode observar, os resultados obtidos são bastante satisfatórios, com especial destaque para o valor da sensibilidade, concluindo-se que, ao aplicar o modelo final a esta amostra independente, há uma percentagem bastante elevada de clientes que entraram incumprimento e foram corretamente classificados como tal. O valor da especificidade, embora inferior, transmite também resultados bastante positivos, sendo, como já referido, numa situação em que não é possível que ambos sejam extremamente elevados, preferível ter uma sensibilidade superior, visto que um falso negativo acaba por ser mais grave no sentido em que o banco está a perder efetivamente dinheiro que emprestou. Isto é claro discutível, uma vez que um falso negativo também é mau no sentido em que o banco poderia estar a gerar lucro com o ganho de juros associados a um empréstimo e acabou por não o conceder devido à errada classificação, mas trata-se de dinheiro que o banco nunca possuiu. O ideal seria ambos os indicadores terem o valor mais elevado possível e a questão de qual das situações é pior depende da perspetiva tomada e é claramente discutível.

Para além disto, observou-se que 70,95% das observações desta amostra obtiveram uma classificação correta, com 29,05% a serem erradamente classificadas. Note-se, porém, que a maioria das classificações erradas diz respeito a clientes que não entraram em incumprimento, sendo, no entanto, classificados como incumpridores segundo o modelo criado, sendo que 28,39% das observações da amostra correspondem a esta situação.

Em suma, a aplicação do modelo final a outra amostra gerou um valor de AUC superior ao do próprio modelo, uma sensibilidade superior, uma especificidade inferior e uma percentagem de classificações corretas inferior. Apesar destas diferenças, se ignorarmos os valores obtidos na criação do modelo, pois já seria de esperar que alguns indicadores se deteriorassem com a sua aplicação a uma amostra que não serviu de base à construção do próprio modelo, os resultados obtidos com a nova amostra são muito satisfatórios e evidenciam a boa qualidade do modelo criado.

## 4.5.Exemplo de Simulação de Perdas

Apenas a título de exemplo, procedeu-se a uma análise bastante superficial da perda a que o banco estaria sujeito caso aceitasse conceder empréstimos às Médias Empresas consoante a classificação gerada pelo modelo final que se obteve.

Para tal, obteve-se uma amostra aleatória relativa a um mês específico do ano, em que cada empresa aparece representada numa só observação.

Aplicando o modelo a esta amostra, obtém-se a matriz de confusão apresentada na Tabela 4.41.

*Tabela 4.41 - Exemplo de Simulação de Perdas: Matriz de Confusão*

| | | **Observados** | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **7** | 49 | 56 |
| | **Negativos** | 3 | **146** | 149 |
| | **Total** | 10 | 195 | 205 |

Obteve-se então uma sensibilidade de 70%, uma especificidade de 74,87% e uma percentagem de empresas corretamente classificadas de 74,63%.

A exposição do banco refletida nesta amostra aleatória distribuir-se-ia como descrito na Tabela 4.42.

*Tabela 4.42 - Exemplo de Simulação de Perdas: Distribuição da Exposição do Banco conforme a Matriz de Confusão Obtida*

| | | **Observados** | | |
|---|---|---|---|---|
| | | **Positivos** | **Negativos** | **Total** |
| **Previstos** | **Positivos** | **5,84 M€** | 38,29 M€ | 44,13 M€ |
| | **Negativos** | 1,50 M€ | **49,01 M€** | 50,51 M€ |
| | **Total** | 7,34 M€ | 87,30 M€ | 94,64 M€ |

Observa-se que a maior parte da exposição (92%) se encontra afeta aos clientes que não entraram em incumprimento nos 12 meses seguintes, o que reflete uma boa gestão por parte do banco.

Assim, caso se aplicasse este modelo à concessão de crédito a Médias Empresas, ter-se ia a seguinte situação:

- O banco sofreria uma perda de 1,50 milhões de euros em clientes a quem se concedeu crédito, por serem classificados como cumpridores, e no futuro entraram em incumprimento, o que corresponde a cerca de 20% da exposição dos clientes incumpridores;
- O banco não teria emprestado 38,29 milhões de euros a clientes classificados como incumpridores segundo o modelo, que, no entanto, não entraram em incumprimento, o que corresponde a 44% da exposição dos clientes cumpridores.

Assim, a perda efetiva do banco seria de 1,5 M€ por incumprimento de crédito, o que corresponde a cerca de 3% da exposição dos clientes a quem o banco teria concedido o empréstimo por serem classificados como cumpridores (50,51 M€). Para além disto, o banco não iria lucrar com os juros associados à exposição de 38,29 M€ dos clientes a quem não seria concedido o empréstimo, mas não entrariam em incumprimento.

É de notar, porém, que, como já referido, esta análise foi realizada apenas a título de exemplo, sendo bastante superficial. Deste modo, sendo que o modelo selecionado evitou a perda de 1,50 M€ em crédito a incumpridores, mas levou à "perda" de dinheiro resultante de juros associados a 38,29 M€ ao não conceder crédito que poderia ter sido bem-sucedido, não é imediatamente conclusivo se a utilização do modelo traz vantagens para o negócio ou não. Para avaliar a situação, seria necessária a realização de cálculos considerando os juros perdidos, o que poderia conduzir, por exemplo, a uma diminuição no ponto de corte final escolhido para o modelo utilizado no âmbito do negócio do banco. No entanto, este tipo de análise não se encontra no âmbito do presente projeto. Em todo o caso, o que importa retirar desta análise é que a utilização do modelo proposto impediu a perda de 5,84 M€ por via do incumprimento de crédito (cerca de 80% da exposição dos clientes que entraram em incumprimento).

# Conclusão

Ao iniciar este projeto, o objetivo era o de criar um ou mais modelos que previssem com qualidade o incumprimento de uma Média Empresa a 12 meses. A criação de mais do que um modelo com qualidade era uma possibilidade, mas não uma certeza, podendo apenas resultar um único modelo.

Com o estudo realizado, percebeu-se que o conjunto de dados disponibilizados se traduziam logo à partida, sem qualquer método de seleção ou transformação de variáveis, num modelo de excelente qualidade, em termos dos vários critérios apresentados, nomeadamente a área abaixo da curva ROC, a sensibilidade, a especificidade e a percentagem de previsões corretas.

Partindo de um modelo completo de grande qualidade, o desafio seria obter um modelo mais simples e de mais fácil aplicação, cuja qualidade não decrescesse significativamente. Foi claro ao longo deste projeto que o objetivo se tornou o de obter um modelo de boa qualidade, mas parcimonioso, visto que se possuía muita informação relativa às Médias Empresas em análise.

Havendo muita informação relativa às empresas presentes na base de dados, haveria sempre a possibilidade de muita dessa informação não ser relevante para prever o seu incumprimento. Desde a aplicação do método de seleção *stepwise* ao modelo inicial completo, após a análise dos fatores de inflação da variância, foi notório que a maioria da informação disponibilizada era de facto relevante, tornando mais desafiante o objetivo de reduzir a dimensionalidade das variáveis do modelo pretendido.

A abordagem selecionada para resolver esta situação foi através da análise dos fatores de inflação da variância, da análise de componentes principais e do método de seleção de variáveis *stepwise*, cuja combinação permitiu reduzir em 55% o número de variáveis iniciais (considerando o conjunto das variáveis *dummy* de cada variável categórica como uma única variável).

Para além da criação do modelo pretendido, com a qualidade desejada, chegou-se ainda a uma outra conclusão que não fazia parte do objetivo original do projeto: foi possível reduzir com sucesso a dimensionalidade das variáveis determinantes para prever o incumprimento a 12 meses de uma Média Empresa sem que o modelo perdesse qualidade de forma significativa, obtendo-se excelentes resultados em qualquer um dos modelos criados neste estudo. Embora a análise dos fatores de inflação da variância tenha sido determinante para obter o modelo final, a metodologia maioritariamente responsável pela redução da dimensionalidade deste modelo foi a análise de componentes principais. Deste modo, concluiu-se que a seleção das componentes principais que explicavam pelo menos 80% da variabilidade das variáveis quantitativas, embora não traduzindo toda a informação que as variáveis originais continham, é uma estratégia que apresenta resultados muito satisfatórios e uma ótima metodologia para obter modelos parcimoniosos, tendo sido a metodologia chave deste projeto.

Em termos da interpretabilidade do modelo, confirmou-se que os fatores que influenciam o incumprimento são, de um modo geral, os empiricamente esperados, como, por exemplo, o facto de a existência de problemas entre sócios aumentar a probabilidade de incumprimento da empresa e o facto de a empresa ter um bom poder de negociação a diminuir.

Por fim, destaca-se a limitação proveniente do facto de o modelo final escolhido não ser de tão fácil interpretação ao nível das variáveis quantitativas, uma vez que estas foram substituídas pelas suas componentes principais. Embora esta não seja uma limitação determinante para a utilização do modelo, é apresentada a alternativa do segundo modelo criado, que apresenta as mesmas variáveis individualmente, tendo uma ótima qualidade. Apesar de existir esta alternativa, ela própria apresenta

também uma limitação por apresentar mais 10 variáveis do que o modelo final selecionado neste projeto. Recorde-se que a comparação entre os dois modelos é apresentada na Tabela 4.38. Deste modo, trata-se de uma questão de preferência do utilizador do modelo e, assim, apresentam-se alternativas de escolha. De notar ainda que, embora se destaque os últimos dois modelos criados, pela maior parcimónia em relação ao modelo completo, o modelo inicial também tem excelente qualidade, o que atesta à qualidade da informação disponibilizada.

Assim, o objetivo do projeto foi cumprido com sucesso, sendo possível aplicar o modelo selecionado para prever a probabilidade de uma Média Empresa entrar em incumprimento no ano seguinte, com base nas suas características atuais, tendo ainda sido obtido um conjunto de alternativas de grande qualidade.

# Referências Bibliográficas


Alpuim, T. (2022). Regressão Logística (Apontamentos de Modelos Lineares). Faculdade de Ciências da Universidade de Lisboa. Obtido em 11 de junho de 2023

Alpuim, T. (2022). Selecção de Variáveis (Apontamentos de Modelos Lineares). Faculdade de Ciências da Universidade de Lisboa. Obtido em 11 de junho de 2023

CFI Team. (20 de fevereiro de 2020). *Credit Risk Analysis Models*. Obtido em 26 de fevereiro de 2023, de Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis-models/

De Laurentis, G., Maino, R., & Molteni, L. (2010). *Developing, Validating and Using Internal Ratings: Methodologies and Case Studies.* John Wiley and Sons Ltd. Obtido em 26 de fevereiro de 2023

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis - 5th Edition.* John Wiley & Sons, Ltd. Obtido em 7 de junho de 2023

Fahrmeir, L., & Kaufmann, H. (março de 1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics 1985, Vol.13, No.1, 342-368*. Obtido em 11 de junho de 2023, de https://projecteuclid.org/journals/annals-of-statistics/volume-13/issue-1/Consistency-and-Asymptotic-Normality-of-the-Maximum-Likelihood-Estimator-in/10.1214/aos/1176346597.full

Fundação Francisco Manuel dos Santos. (23 de fevereiro de 2023). *Endividamento das sociedades não financeiras privadas: total e por dimensão*. Obtido em 27 de maio de 2023, de PORDATA: https://www.pordata.pt/portugal/endividamento+das+sociedades+nao+financeiras+privadas+total+e+por+dimensao-2924

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Gastos com pessoal das pequenas e médias empresas: total e por dimensão*. Obtido em 27 de maio de 2023, de PORDATA: https://www.pordata.pt/Portugal/Gastos+com+pessoal+das+pequenas+e+m%C3%A9dias+empresas+total+e+por+dimens%C3%A3o-2935

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Pequenas e médias empresas em % do total de empresas: total e por dimensão*. Obtido em 28 de maio de 2023, de PORDATA: https://www.pordata.pt/portugal/pequenas+e+medias+empresas+em+percentagem+do+total+de+empresas+total+e+por+dimensao-2859

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Pequenas e médias empresas: total e por dimensão*. Obtido em 27 de maio de 2023, de PORDATA: https://www.pordata.pt/portugal/pequenas+e+medias+empresas+total+e+por+dimensao-2927

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Pessoal ao serviço nas empresas: total e por setor de atividade económica*. Obtido em 21 de agosto de 2023, de PORDATA: https://www.pordata.pt/portugal/pessoal+ao+servico+nas+empresas+total+e+por+setor+de+atividade+economica-2895

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Pessoal ao serviço nas pequenas e médias empresas*. Obtido em 27 de maio de 2023, de PORDATA:

https://www.pordata.pt/portugal/pessoal+ao+servico+nas+pequenas+e+medias+empresas-2931

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Volume de negócios das empresas: total e por setor de atividade económica*. Obtido em 21 de agosto de 2023, de PORDATA: https://www.pordata.pt/portugal/volume+de+negocios+das+empresas+total+e+por+setor+de+atividade+economica-2913

Fundação Francisco Manuel dos Santos. (28 de fevereiro de 2023). *Volume de negócios das pequenas e médias empresas: total e por dimensão*. Obtido em 27 de maio de 2023, de PORDATA: https://www.pordata.pt/portugal/volume+de+negocios+das+pequenas+e+medias+empresas+total+e+por+dimensao-2932

Gomes, J. J. (2021). Regressão Linear. Faculdade de Ciências da Universidade de Lisboa. Obtido em 6 de junho de 2023

Gonçalves, C., Santos, D., Rodrigo, J., & Fernandes, S. (2020). *Contabilidade Financeira Explicada - 4ª edição.* Porto: Vida Económica. Obtido em 3 de julho de 2023

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression - Second Edition.* John Wiley & Sons, Inc. Obtido em 15 de junho de 2023

Jacquinet, M. (2 de outubro de 2019). Introdução breve ao conceito de Cash Flow (ou fluxo de caixa). Universidade Aberta. Obtido em 3 de junho de 2023, de https://repositorioaberto.uab.pt/handle/10400.2/8613

Martins, M. E. (2014). Coeficiente de correlação amostral. *Revista de Ciência Elementar, Volume 2, Número 2*. Obtido em 6 de junho de 2023, de https://rce.casadasciencias.org/rceapp/art/2014/042/

*Mútuo - Lexionário*. (s.d.). Obtido em 25 de fevereiro de 2023, de Diário da República: https://diariodarepublica.pt/dr/lexionario/termo/mutuo

Pariente, R. (22 de maio de 2018). *The bank's core business*. Obtido em 25 de fevereiro de 2023, de BBVA: https://www.bbva.com/en/banks-core-business/

*PME Pequenas e Médias Empresas*. (2 de novembro de 2021). Obtido em 27 de maio de 2023, de Portugal 2020: https://portugal2020.pt/glossario/pme-pequenas-e-medias-empresas/

Quintaneiro, J., & Martins, B. (maio de 2007). Demonstração de Resultados (DR). Instituto Politécnico de Coimbra. Obtido em 27 de maio de 2023, de https://associativismo.cm-vfxira.pt/images/stories/formacao/formacoes/2013/formacao_2013_doc2.2.pdf

Rajani, A. (16 de dezembro de 2020). *Classifications and Key Concepts of Credit Risk (I)*. Obtido em 26 de fevereiro de 2023, de Linkedin: https://www.linkedin.com/pulse/classifications-key-concepts-credit-risk-i-asif-rajani/

Rajani, A. (14 de dezembro de 2020). *The Credit Decision*. Obtido em 26 de fevereiro de 2023, de Linkedin: https://www.linkedin.com/pulse/credit-decision-asif-rajani/

REGULAMENTO (UE) N. o 575/2013 DO PARLAMENTO EUROPEU E DO CONSELHO de 26 de junho de 2013 relativo aos requisitos prudenciais das instituições de crédito e que altera o Regulamento (UE) n. o 648/2012 (Texto relevante para efeitos do EEE). (30 de setembro de

2021). Portugal. Obtido em 27 de maio de 2023, de https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02013R0575-20210930

Reis, T. (12 de setembro de 2019). *Entendendo o Modelo de Negócio dos Bancos*. Obtido em 25 de fevereiro de 2023, de Investing.com: https://br.investing.com/analysis/entendendo-o-modelo-de-negocio-dos-bancos-200431355

Severino, E. (2020). TESTES DO QUI-QUADRADO, PARA A INDEPENDÊNCIA E PARA A HOMOGENEIDADE, EM TABELAS DE CONTINGÊNCIA (Análise de Dados 20/21). Faculdade de Ciências da Universidade de Lisboa.

Telhada, J., & Fonseca, R. (2021). Gestão Financeira 2021/22 - Análise Financeira (AF). Faculdade de Ciências da Universidade de Lisboa.

Telhada, J., & Fonseca, R. (2021). Gestão Financeira 2021/22 - Contabilidade financeira (CoF). Faculdade de Ciências da Universidade de Lisboa.

Telhada, J., & Fonseca, R. (2021). Gestão Financeira 2021/22 - Estrutura de capital da empresa (ECE). Faculdade de Ciências da Universidade de Lisboa.

The Investopedia Team. (15 de março de 2022). *Credit Risk: Definition, Role of Ratings, and Examples*. Obtido em 26 de fevereiro de 2023, de Investopedia: https://www.investopedia.com/terms/c/creditrisk.asp

Varella, C. A. (11 de dezembro de 2008). Análise de Componentes Principais. Universidade Federal Rural do Rio de Janeiro. Obtido em 7 de junho de 2023, de http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf