INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Modelos preditivos de insolvências e falências em empresas portuguesas: o impacto de indicadores financeiros e não financeiros

Mariana Viegas da Silva Ildefonso

Mestrado em Contabilidade

Orientadores:
Professor Raul Manuel da Silva Laureano, Professor Associado, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Professor Miklos A. Vasarhelyi, Professor reconhecido pela KPMG em Sistemas de Informação Contabilística e Diretor do Centro de investigação de Contabilidade de *Rutgers* e do *Continuous Auditing & Reporting Lab*

Setembro, 2023

Departamento de Contabilidade

Modelos preditivos de insolvências e falências em empresas portuguesas: o impacto de indicadores financeiros e não financeiros

Mariana Viegas da Silva Ildefonso

Mestrado em Contabilidade

Orientadores:
Professor Raul Manuel da Silva Laureano, Professor Associado, ISCTE Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Professor Miklos A. Vasarhelyi, Professor reconhecido pela KPMG em Sistemas de Informação Contabilística e Diretor do Centro de investigação de Contabilidade de *Rutgers* e do *Continuous Auditing & Reporting Lab*

Setembro, 2023

Agradecimentos

Este ano da execução da dissertação de mestrado foi um ano particularmente difícil. Ainda assim, tive a sorte de ter sempre pessoas que me apoiaram, me deram força e me ajudaram no que eu mais precisava, sem que nunca me deixassem desistir.

Ao professor Raul Laureano agradeço por toda a paciência, por todo o tempo investido em mim e na minha investigação, por toda a ajuda despendida e por toda a força para que nunca desistisse de entregar o trabalho.

Ao professor Miklos Vasarhelyi agradeço pelo tempo despendido, pela paciência e pelas críticas construtivas para que este trabalho corresse da melhor forma possível.

Aos meus pais, Luís e Sara, e madrasta, Leonor, por todo o apoio que me deram porque sem eles nada disto tinha sido possível, por todos os momentos bons com que me prendaram e por todos os conselhos que me deram.

Aos meus avós, Carmen, Francisco, Maria, Olívia, Otília e Rodrigo, por terem sido sempre o meu escape quando os tempos mais difíceis se avizinhavam e por terem sido sempre o abraço e força que precisava.

Aos meus irmãos, Joana, Helena, Rodrigo e Vanda por terem respeitado o tempo que não estive com eles e por me terem sempre apoiado independentemente das circunstâncias.

Aos meus tios, Clara, Marta e Tó, e primos, Beatriz, Helena, Henrique, João Afonso, Pedro, Raquel e Rita, por terem sido sempre parte das minhas gargalhadas e das conversas de longas horas.

Aos meus amigos, Daniela, Inês, Luís, Marta, em especial à Andreia, por tanto que aturou nesta fase final da dissertação de mestrado, pelas risadas sem mais e pela paciência, à Catarina, por toda a companhia que me fez ao longo da execução da dissertação de mestrado, por toda a ajuda que me deu, paciência e, acima de tudo, por ter estado sempre aqui, e à Rita, pela companhia que me fez e pela motivação que me deu.

Aos meus vizinhos, Bruno, Daniela, Licínia, Paulo, Rosa, Sofia, Tomás e Valério, por todos os momentos de brincadeira para que pudesse desanuviar a cabeça e voltar com maior motivação.

Ao meu tio Tiago e ao meu bisavô Luís, que de longe me têm sempre acompanhado, obrigada por estarem sempre tão presentes.

A todos os que estiveram sempre comigo nesta caminhada, nunca me esquecerei.

Muito obrigada,

Mariana

Resumo

Em Portugal, desde a crise financeira de 2008, o número de empresas que entram em insolvência/falência é bastante elevado e preocupante pelos impactos que causam na economia e na sociedade. Apesar de já existirem diversos modelos preditivos das insolvências e falências, cujos preditores são, essencialmente, indicadores financeiros, a previsão de insolvências e falências ainda é crítica nos dias de hoje, pelo que é importante continuar a investigar e a criar modelos com maior precisão que os anteriores. Deste modo, o presente estudo avalia o impacto de indicadores financeiros e não financeiros na predição das insolvências e falências. Para tal, recorre-se a técnicas preditivas de análise de dados mais avançadas, nomeadamente árvores de decisão com os algoritmos CART, CHAID e C5.0, de forma a analisar o impacto dos indicadores entre os anos de 2013 e 2022, de uma amostra de 707.291 empresas, 642.983 ativas e 64.308 insolventes/falidas. Os resultados obtidos permitem identificar uma relação entre os indicadores financeiros e a insolvência/falência das empresas, prevendo-se uma percentagem de empresas corretamente classificadas de 82%. O principal contributo desta investigação é gerar conhecimento sobre a inviabilidade das empresas através dos indicadores financeiros e não financeiros, recorrendo-se a técnicas nunca utilizadas em modelos preditivos aplicados a Portugal.

Palavras-chave: Insolvência, Falência, Árvores de Decisão, Previsão

JEL Classification System: M10, M41

Abstract

In Portugal, since the financial crisis of 2008, the number of companies entering insolvency/bankruptcy is very high and worrying due to the impact they have on the economy and society. Although there are already several predictive models for insolvencies and bankruptcies, whose predictors are essentially financial indicators, the prediction of insolvencies and bankruptcies is still critical today, so it is important to continue researching and creating models with greater precision than the previous ones. Therefore, this study assesses the impact of financial and non-financial indicators in predicting insolvencies and bankruptcies. To this end, more advanced data analysis techniques were used, namely decision trees with the CART, CHAID and C5.0 algorithms, in order to analyze the impact of the indicators between 2013 and 2022, from a sample of 707,291 companies, 642,983 active and 64,308 insolvent/bankrupt. The results obtained show a relationship between financial indicators and company insolvency/bankruptcy, with a predicted percentage of correctly classified examples of 82%. The main contribution of this research is to generate knowledge about the viability of companies through financial and non-financial indicators, using techniques never used before in predictive models applied to Portugal.

Keywords: Insolvency, Bankruptcy, Decision Trees, Prediction

JEL Classification System: M10, M41

Lista de Abreviaturas, Acrónimos e Siglas

AML	Área Metropolitana de Lisboa
AUC	*Area Under the Receiver Operating Characteristic Curve*
CART	*Classification and Regression Trees*
CHAID	*Chi-squared Automatic Interaction Detection*
CIRE	Código da Insolvência e da Recuperação de Empresas
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CSC	Código das Sociedades Comerciais
EUA	Estados Unidos da América
e.g.	Em geral (do latim, *exempli gratia*)
i.e.	Isto é
Lda	Sociedades por quotas de responsabilidade limitada
NACE	Nomenclatura Estatística das Atividades Económicas da Comunidade Europeia
PCCC	Percentagem de casos corretamente classificados
PIB	Produto Interno Bruto
PME	Pequenas e Médias Empresas
RL	Revisão de Literatura
RSL	Revisão Sistemática de Literatura
VC	V de Cramer
WoS	*Web of Science*

Índice Geral

Índice de Figuras

Índice de Tabelas

1. Introdução

Esta introdução apresenta o tema a ser investigado, bem como os conceitos de dificuldades financeiras, insolvências e falências, realçando a importância da sua previsão em todo o mundo. Adicionalmente, expõe o problema e a questão de investigação, os objetivos e a abordagem metodológica desenvolvida, e, termina, com a estrutura e organização da dissertação.

1.1. Tema e a sua delimitação

A previsão de dificuldades financeiras, insolvências e falências é cada vez mais importante no que toca a investidores, credores, bancos, entre outros *stakeholders* das empresas, devido à probabilidade de incumprimento para com os mesmos (Barboza *et al.*, 2017).

Desde a crise financeira de 2008 a gestão de risco de crédito tem sido uma prioridade para as empresas, levando, assim, à realização de estudos visando criar o melhor modelo preditivo para estas condições (Barboza *et al.*, 2017). As primeiras descobertas de modelos preditivos foram feitas por Beaver (1966), com a utilização de indicadores financeiros. Contudo, indicadores não financeiros relacionados com a gestão também se mostram importantes na previsão de insolvências (Chen e Du, 2009; Cooper e Uzun, 2019). Assim, este estudo inova com a complementação de indicadores não financeiros na previsão de insolvências tendo em conta modelos já utilizados na literatura com variáveis financeiras e aplicando-se o estudo a Portugal.

De acordo com Isayas (2021), uma empresa encontra-se em dificuldades financeiras quando não é capaz de cumprir as suas obrigações financeiras para com os credores, o que, em contrapartida, leva a uma reestruturação ou falência da empresa. Segundo o artigo 1º do Código da Insolvência e da Recuperação de Empresas (CIRE) "o processo de insolvência é um processo de execução universal que tem como finalidade a satisfação dos credores pela forma prevista num plano de insolvência, baseado, nomeadamente, na recuperação da empresa compreendida na massa insolvente, ou, quando tal não se afigure possível, na liquidação do património do devedor insolvente e a repartição do produto obtido pelos credores". Sendo assim, uma empresa pode encontrar-se no processo de falência quando a mesma está num período de crise e se depara com muitas dificuldades financeiras que a impedem de cumprir as suas obrigações, tornando-as economicamente inviáveis (Eportugal, 2023).

Enquanto a previsão de falências se concentra principalmente na previsão do fim do ciclo de vida de uma empresa, com relativamente poucas hipóteses de sobrevivência através de

reestruturação, a previsão de dificuldades financeiras é uma ocorrência mais comum, quando um negócio tem dificuldades temporárias no cumprimento das suas obrigações. A previsão das dificuldades financeiras das empresas desempenha um papel cada vez mais importante na sociedade de hoje, uma vez que tem um impacto significativo nas decisões de empréstimo e na rentabilidade das instituições financeiras. Desde a crise financeira ocorrida em 2008, houve um maior nível da atenção por parte das instituições financeiras relativamente ao risco de crédito e à previsão de dificuldades financeiras das empresas. Daqui advém então a importância de cada empresa perceber o tipo de cliente que tem à sua mercê, de forma a evitar falta de pagamentos, por exemplo. Com esta crise financeira, as instituições financeiras começaram a travar a concessão de financiamentos e as empresas começaram a ressentir-se com o elevado nível de endividamento a que não conseguiam fazer face, em simultâneo, com a recusa de acesso ao crédito por parte das instituições financeiras (Varejão, J., 2020).

Em Portugal, a crise de 2008 teve como principais indicadores o Produto Interno Bruto (PIB) per capita, que teve uma queda de 4,4%; a taxa de desemprego, que subiu ligeiramente (isto é, (i.e.) 1,4%) porém não cessou mesmo após este abalo financeiro; e o indicador do sentimento económico (i.e., indicador calculado pela Comissão Europeia que mede a confiança e as expectativas quanto à economia de consumidores e empresas europeias), que teve um decréscimo em 24,2 pontos percentuais (Varejão, 2020).

Figura 1.1: Evolução de dificuldades financeiras a insolvência e falência

Fonte: Elaboração própria

1.2. Problema e questão de investigação

A previsão da entrada em processos de insolvência de uma empresa é importante pois prepara os gestores para os desafios que se avizinham. Para os investidores é uma forma de perceberem se investir numa certa empresa é viável ou não tendo em conta a sua saúde financeira.

De acordo com Simić *et al.* (2012), a previsão da falência das empresas é crucial para a prevenção ou mitigação de ciclos económicos negativos na economia, permitindo identificar

futuras falhas e fornecendo atempadamente chamadas de atenção para futuras dificuldades financeiras. Ao longo dos últimos anos tem havido um crescimento moderado de insolvências/falências, tendo sido notório um crescimento no ano de 2020 devido à pandemia do COVID-19 e com a consequência da paragem momentânea das economias mundiais (Tabela 1.1).

Tabela 1.1: Evolução dos processos de insolvência e falências em Portugal

	Número	**Taxa de variação**
2017	2 658	-
2018	2 332	-12,26%
2019	2 135	-8,45%
2020	**2 183**	**2,25%**
2021	1 930	-11,59%
2022	1 598	-17,20%

Fonte: INE (2023)

Apesar de atualmente existirem diversos modelos preditivos de dificuldades financeiras, falências e insolvências, ainda há muita informação por descobrir e limitações a serem colmatadas, para que possam ser criados modelos mais precisos e com uma maior importância para indicadores não financeiros. Com esse intuito, este estudo visa responder à seguinte questão de investigação: "De que modo a estrutura financeira e societária de uma empresa permite prever a sua entrada em processos de insolvência e falências?".

1.3. Objetivos e contributos

Tendo em conta a questão de investigação formulada anteriormente, o objetivo geral desta investigação é identificar, através da criação de um modelo preditivo, qual o impacto da estrutura financeira e societária de uma empresa na compreensão das insolvências/falências das mesmas. Para alcançar este objetivo geral são definidos três objetivos específicos: (1) caracterizar a situação financeira e não financeira de empresas insolventes/falidas[1], e ativas; (2) criar o modelo preditivo das insolvências/falências; e (3) identificar perfis de empresas não ativas e ativas.

[1] A partir deste momento, as empresas insolventes e falidas passam a denominar-se empresas não ativas para facilidade na leitura.

Com a concretização dos objetivos e, consequente, resposta à questão de investigação, este estudo contribui para o desenvolvimento dos modelos preditivos das insolvências/falências, com a inclusão de indicadores não financeiros. Com a conceção deste modelo preditivo podem ser identificados perfis de empresas propensas a entrar em insolvência/falência e, como tal, os gestores das empresas conseguem, atempadamente, delinear estratégias que as possam evitar (Park *et al.*, 2021).

1.4. Abordagem Metodológica

Atendendo à questão de investigação e aos objetivos propostos, esta investigação adota uma metodologia quantitativa (Major, 2017), ao analisar os dados de uma amostra de 707 291 empresas, 64 308 com *status* não ativa e 642 983 com *status* ativa.

Em termos metodológicos recorre-se à metodologia CRISP-DM visto ser o tipo de processo mais utilizado entre estudos do género (Lebkiri *et al.*, 2021).

Tendo em conta os objetivos propostos, de forma a garantir a qualidade do estudo, recorre-se a uma técnica já utilizada na literatura e que demonstra bons resultados, árvores de decisão, para construção do modelo preditivo (em geral (e.g.), Olson *et al.*, 2012).

1.5. Estrutura e organização da dissertação

No que respeita à estrutura da dissertação e, de forma a compreender o impacto dos indicadores financeiros e não financeiros na predição das insolvências/falências, esta organiza-se em cinco capítulos.

No primeiro capítulo, são abordados o tema, a relevância do mesmo, o problema e questão de investigação, bem como apresentados os objetivos e contributos da investigação. Adicionalmente, é apresentada a abordagem metodológica adotada para a concretização desses mesmos objetivos. O segundo capítulo, o da revisão de literatura, em que se aborda a problemática das dificuldades financeiras, insolvências e falências, e os seus impactos, e os modelos preditivos já desenvolvidos noutros estudos. O terceiro capítulo, descreve a metodologia adotada e todo o procedimento de recolha e análise de dados. Os resultados e discussão dos mesmos são apresentados no capítulo quatro. Por fim, no capítulo cinco são apresentadas as conclusões, destacando-se os contributos, as limitações e as pistas futuras para outras investigações.

2. Revisão de Literatura

Este capítulo procura sistematizar o conhecimento científico sobre a entrada em processos de insolvência de diferentes empresas, falências e dificuldades financeiras mais concretamente a sua previsão a partir de técnicas, ditas, de *machine learning* e de *data mining*.

As técnicas de *machine learning* definem-se, de acordo com Mizik e Hanssens (2018), como o estudo de métodos ou algoritmos criados para aprenderem os padrões dos dados recolhidos e conseguirem obter previsões com base nesses padrões. Assim sendo, uma característica fundamental desta técnica reside na sua capacidade de realizar previsões precisas para dados que não fazem parte da amostra original, graças à sua habilidade em aprender os padrões a partir dos dados da amostra. Esta técnica tem duas ramificações: (1) aprendizagem supervisionada e (2) aprendizagem não supervisionada. Os mesmos autores definem a aprendizagem supervisionada como um processo em que é previsto um valor de alguma variável alvo tendo em conta diversas variáveis independentes; em contrapartida, a aprendizagem não supervisionada não tem nenhuma variável alvo pelo que o objetivo não é prever nenhum valor, mas encontrar grupos com características semelhantes e que se relacionem entre si. As técnicas de *data mining*, de acordo com Ma *et al.* (2022), são técnicas que utilizam algoritmos matemáticos para encontrar informação oculta numa enorme quantidade de dados e analisarem os padrões indexados a estes dados.

De acordo com Ahsan e Siddique (2022), a RSL é "*uma revisão onde são formuladas perguntas e em que são utilizados procedimentos sistemáticos e explícitos para selecionar e fazer uma avaliação crítica da investigação relevante, a fim de recolher e avaliar os dados dos estudos incluídos na revisão*" [2]. Esta envolve três passos principais: (1) a determinação da questão de investigação; (2) elaboração da estratégia de pesquisa e seleção; e, por fim, (3) a realização da extração e síntese dos dados (Putrada *et al.,* 2022).

No subcapítulo seguinte é apresentado o protocolo da RSL, bem como o processo de recolha dos artigos analisados. Os artigos são apresentados e analisados, de forma sistemática, os resultados que permitem responder às questões definidas no protocolo, incluindo, a identificação das diferentes técnicas de predição de insolvência, falência e dificuldades

[2] Tradução livre de "*A systematic literature review (SLR) is a review in which questions are formulated and systematic and explicit procedures are used to find select, and critically appraise relevant research in order to gather and evaluate data from the studies included in the review.*"

financeiras. Por fim, os artigos utilizados na RSL são alvo de avaliação da sua qualidade recorrendo a diferentes critérios de avaliação, igualmente, identificados no protocolo.

Após apresentados os conceitos relevantes ao estudo, segue-se a revisão sistemática da literatura (RSL) dos artigos selecionados, que se encontra dividida no protocolo e na análise dos artigos.

2.1. Revisão sistemática da literatura sobre predição de falências, insolvências e dificuldades financeiras

2.1.1. Protocolo para a revisão sistemática

Atendendo ao objetivo da investigação, esta revisão da literatura propõe-se a responder à pergunta: "De que modo tem sido efetuada a previsão de entrada em insolvência/falência/dificuldades financeiras de uma empresa ou o aparecimento de dificuldades financeiras na mesma?". Mais especificamente, visa responder às três questões seguintes: (1) "Quais os âmbitos e contextos em que foram realizados os estudos?"; (2) "Quais os modelos utilizados na previsão das insolvências/falências/dificuldades financeiras e sua validação?"; e (3) "Como são avaliados os resultados dos estudos?".

Os artigos incluídos nesta RSL são selecionados a partir de uma base de dados com diferentes publicações científicas, tendo em conta as palavras-chave, resumo, título e critérios de inclusão/exclusão. É selecionada a plataforma *Web of Science* (*WoS*) (https://www.webofscience.com/) devido à sua relevância em áreas de investigação como a gestão e a economia, tendo bastantes *journals* indexados, e pelo facto de ser uma plataforma mais seletiva e com informação mais consistente, como referido por Esteban *et al.* (2022).

A procura automática de artigos na base de dados é elaborada a partir da aplicação de uma *query* ao título, resumo e palavras-chave do artigo. As palavras-chave utilizadas na *query* resultam da identificação na literatura de termos identificativos de insolvência, falência e dificuldades financeiras, de previsão e de métodos de análise de dados, tendo sido validados por especialistas (e.g., professores de contabilidade e finanças). Assim, a *query* aplicada é "*(predict* OR detect*) AND (bankrupt* OR "financial distress" OR insolven*) AND ("machine learning" OR "data mining")*". A partir daqui foram selecionados 42 artigos, que posteriormente serão submetidos a critérios de inclusão e exclusão, conforme apresentado na Tabela 2.1.

Tabela 2.1: Critérios de inclusão e exclusão

Critérios de inclusão	Artigos que abordam insolvência e falência em termos de sua previsão
	Artigos que analisam indicadores financeiros das empresas
	Artigos com componente empírica
Critérios de exclusão	*Proceeding papers, Early Acess, Review Articles*
	Artigos que não tenham fator de impacto

Fonte: Elaboração própria

Tendo em conta os critérios de inclusão e exclusão, bem como da leitura do resumo de cada artigo, excluem-se 21 artigos, estando o procedimento apresentado resumidamente na

Figura *2.1*. Assim sendo, são incluídos na revisão da literatura (RL) desta investigação 21 artigos, apresentados na Tabela 2.2.

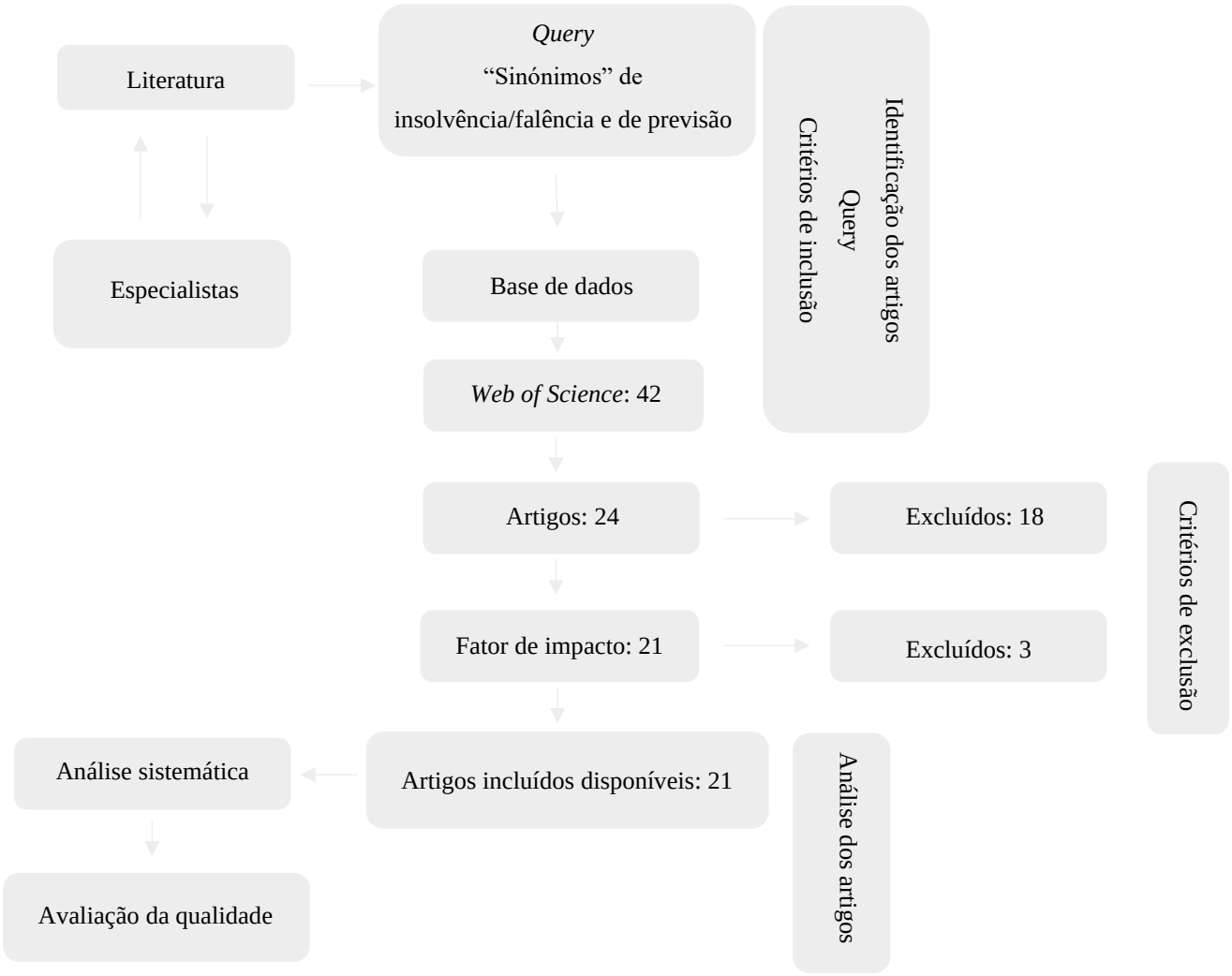

Figura 2.1: Processo de seleção dos artigos para revisão sistemática da literatura

Fonte: Elaboração própria

A análise da Tabela 2.2 permite concluir que apenas três *journals* (i.e., *Journal of Forecasting*; *Knowledge-Based Systems*; *Expert Systems with Applications)* publicaram dois artigos, tendo todos os outros apenas um artigo nesta amostra. Importa ainda referir que nenhum *journal* é da área da contabilidade, sendo apenas um da área de finanças (*Quantitative Finance*) e que nenhum autor surge em mais do que um artigo.

Tabela 2.2: Artigos incluídos na RSL

ID	Ano	Título	*Journal*[3]	Autores
1	2022	*Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data*	*Computational Economics*	Kim, H.; Cho, H.; Ryu, D.
2	2022	*A comparison of static, dynamic and machine learning models in predicting the financial distress of chinese firms*	*Romanian Journal of Economic Forecasting*	Bin Yousaf, U.; Jebran, K.; Wang, M.
3	2022	*Machine-learning models for bankruptcy prediction: do industrial variables matter?*	*Spatial Economic Analysis*	Bragoli, D.; Ferretti, C.; Ganugi, P.; Marseguerra, G.; Mezzogori, D.; Zammori, F.
4	2021	*Predicting bankruptcy of local government: A machine learning approach*	*Journal of Economic Behavior & Organization*	Antulov-Fantulin, N.; Lagravinese, R.; Resce, G.
5	2021	*Explainability of Machine Learning Models for Bankruptcy Prediction*	*Ieee Access*	Park, M.; Son, H.; Hyun, C.; Hwang, H.
6	2021	*Financial Distress Prediction for Small and Medium Enterprises Using Machine Learning Techniques*	*Inzinerine Ekonomika-Engineering Economics*	Malakauskas, A.; Lakstutiene, A.
7	2020	*Predicting the Insolvency of SMEs Using Technological Feasibility Assessment Information and Data Mining Techniques*	*Sustainability*	Lee, S.; Choi, K.; Yoo, D.
8	2020	*Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods*	*Journal of Forecasting*	Tang, X.; Li, S.; Tan, M.; Shi, W.
9	2020	*Predicting bank insolvencies using machine learning techniques*	*International Journal of Forecasting*	Petropoulos, A.; Siakoulis, V.; Stavroulakis, E.; Vlachogiannakis, N.
10	2019	*Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design*	*Quantitative Finance*	Lahmiri, S.; Bekiros, S.
11	2019	*A new perspective of performance comparison among machine learning algorithms for financial distress prediction*	*Applied Soft Computing*	Huang, Y.; Yen, M.
12	2017	*Machine learning models and bankruptcy prediction*	*Expert Systems with Applications*	Barboza, F.; Kimura, H.; Altman, E.
13	2015	*The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches*	*Knowledge-Based Systems*	Zhou, L.; Lu, D.; Fujita, H.
14	2015	*Prediction of financial distress: An empirical study of listed Chinese companies using data mining*	*European Journal of Operational Research*	Geng, R.; Bose, I.; Chen, X.
15	2012	*Comparative analysis of data mining methods for bankruptcy prediction*	*Decision Support Systems*	Olson, D.; Delen, D.; Meng, Y.
16	2012	*A Robust Data-Mining Approach to Bankruptcy Prediction*	*Journal of Forecasting*	Divsalar, M.; Roodsaz, H.; Vahdatinia, F.; Norouzzadeh, G.; Behrooz, A.
17	2009	*Using neural networks and data mining techniques for the financial distress prediction model*	*Expert Systems with Applications*	Chen, W.; Du, Y.
18	2008	*Data mining method for listed companies' financial distress prediction*	*Knowledge-Based Systems*	Sun, J.; Li, H.
19	2004	*Multiple criteria linear programming data mining approach: An application for bankruptcy prediction*	*Data Mining and Knowledge Management*	Kwak, W.; Shi, Y.; Cheh, J.; Lee, H.
20	2004	*Variable selection in data mining: Building a predictive model for bankruptcy*	*Journal of the American Statistical Association*	Foster, D.; Stine, R.
21	2000	*Extracting predictors of corporate bankruptcy: Empirical study on data mining methods*	*Knowledge Discovery and Data Mining, Proceedings: Current Issues and New Applications*	Shirata, C.; Terano, T.

Fonte: Elaboração própria

[3] Ver a avaliação de cada *journal no* Anexo *D – Avaliação dos* journals pelo Web of Science

Após a leitura e revisão dos artigos selecionados, é feita uma avaliação da qualidade dos mesmos tendo em conta a sua contribuição para a concretização do objetivo e, em particular, para a resposta às três questões colocadas, apresentadas na Tabela 2.3. Esta avaliação tem por base 13 itens, colocados na forma de questão, sendo que a cada é atribuída uma cotação de 0 se não responde à questão; 0,5 se responde parcialmente e 1 se responde totalmente.

Tabela 2.3: Critério de qualidade para a avaliação dos artigos

Âmbitos e contextos da realização dos estudos	Q1.1: Define claramente insolvência/falência/dificuldade financeira?
	Q1.2: É evidenciada a diferença entre insolvências, falências e dificuldades financeiras?
	Q1.3: Justifica a finalidade e o contexto do estudo?
	Q1.4: A amostra é caracterizada adequadamente?
	Q1.5: Apresenta a metodologia adequadamente?
Modelos utilizados na previsão das falências/insolvências/dificuldades financeiras e sua validação	Q2.1: Apresenta, justificando adequadamente, as variáveis utilizadas?
	Q2.2: Apresenta, descreve e justifica a(s) técnica(s)/ algoritmo(s) utilizada/o(s)?
	Q2.3: Compara diferentes modelos preditivos?
	Q2.4: Avalia convenientemente o(s) modelo(s)?
	Q2.5: São realizadas técnicas de robustez e de sensibilidade?
Avaliação dos estudos	Q3.1: Os resultados do estudo são devidamente discutidos?
	Q3.2: São identificados claramente os contributos (teóricos e práticos) e os impactos do estudo?
	Q3.3: São identificadas claramente as limitações do estudo?

Fonte: Elaboração própria

2.1.2. Análise sistemática dos artigos

Os 21 artigos selecionados para análise, de acordo com o protocolo, são agora analisados em três diferentes vertentes. Primeiro, caracterizam-se os âmbitos e objetivos dos estudos, referindo-se ainda o tipo de empresa estudada. Depois, analisam-se os modelos utilizados nas diferentes previsões tendo em conta o objetivo do estudo, caracterizando-se também a sua forma de validação. Finalmente, passa-se para a fase de avaliação que também esta se encontra dividida em duas fases: primeira em que se avaliam os resultados dos estudos e, seguida pela avaliação da qualidade dos artigos elegidos.

2.1.2.1. Âmbitos e contextos

Os artigos selecionados descrevem estudos muito diversificados, quer quanto ao seu âmbito e objetivo (Tabela 2.4), quer quanto aos tipos de empresas estudados (Tabela 2.5).

Relativamente ao âmbito verifica-se que mais de metade dos artigos selecionados, 11 artigos do total de 21, abordam o tema das falências, sendo que as insolvências são o foco de apenas dois estudos (Lee *et al.,* 2020; Petropoulos *et al.,* 2020). Destaca-se, ainda, que a previsão de dificuldades financeiras surge em oito estudos, o que revela alguma preocupação dos investigadores em prever com antecedência possíveis falências, uma vez que as dificuldades financeiras são os primeiros sinais de que a sustentabilidade financeira das empresas possa estar comprometida (e.g., Tang *et al.,* 2020).

Naturalmente, todos os estudos incidem sobre modelos de previsão. No entanto, alguns destes não têm como principal objetivo criar modelos preditivos das falências, insolvências ou dificuldades financeiras, mas aplicar novos algoritmos para problemas de previsão (e.g., algoritmos de *machine learning*) para ver de que forma estes permitem melhorar a qualidade das mesmas (e.g., Barboza *et al.,* 2017; Kim *et al.,* 2022). Contudo, outros estudos pretendem elaborar de raiz o melhor modelo preditivo de falências, insolvências ou dificuldades financeiras (e.g., Divsalar *et al.,* 2012; Sun e Li, 2008).

Já no que respeita aos períodos (temporais) da análise, constata-se igualmente muita diversidade uma vez que este varia entre três meses (Huang e Yen, 2019) e 29 anos (Barboza *et al.,* 2017). Assim, muitos estudos consideram nos seus modelos possíveis impactos de crises, como por exemplo a tão conhecida crise financeira de 2008 (e.g., Barboza *et al.,* 2017; Petropoulos *et al.,* 2020).

Tabela 2.4: Contextualização dos artigos da RSL– Âmbito/objetivos dos estudos

ID	Âmbito	Objetivo do estudo	Período de estudo	Anos da recolha de dados
1	Falências	Perceber se as previsões de falência de empresas podem ser melhoradas utilizando os algoritmos *RNN* e *LSTM*	13 anos	2007-2019
2	DF	Comparar o desempenho dos modelos estáticos, dinâmicos e de *ML* na previsão de DF nas empresas cotadas	3 anos-27 anos	1992-2018
3	Falências	Dar ênfase à importância de variáveis industriais/regionais na previsão de falências	1 ano	2007-2015
4	Falências	Enriquecer a literatura sobre *EWM*, preenchendo esta lacuna através da utilização de metodologias de *ML* de previsão de falências	1 ano	2009-2016
5	Falências	Mostrar que a importância das características medida através do *LIME* pode ser uma generalização consistente da importância medida a partir dos modelos baseados em árvores	NE	2009-2015
6	DF	Propor um modelo para prever DF em PME que inclua fatores temporais e que preveja os *ratings* dos créditos das PME num horizonte temporal de 1 ano	2 anos	NE
7	Insolvências	Comparar os modelos de previsão de insolvências em PME e encontrar o que tenha maior *accuracy*	5 anos	2010-2014
8	DF	Explorar os fatores integrados e os múltiplos modelos que podem melhorar o desempenho dos modelos para a previsão de DF em empresas chinesas cotadas	3 anos	2013-2018
9	Insolvências	Comparar diferentes algoritmos na previsão de insolvências bancárias	7 anos	2008-2014
10	Falências	Prever falências utilizando técnicas de *mimetic intelligent emulating*	NE	NE
11	DF	Comparar a performance de modelos de *ML* na previsão de DF em empresas cotadas	3 meses	2010-2016
12	Falências	Melhoria na precisão da previsão de falências, comparando modelos e usando técnicas de *ML*	29 anos	1985-2013
13	DF	Investigar o desempenho de diferentes modelos de previsão de DF em empresas chinesas com abordagens de *feature selection* baseadas em conhecimentos de *data mining*	2 anos	2001-2011
14	DF	Aplicar *data mining* para prever quais as empresas cotadas suscetíveis de receber o rótulo *ST*	3 anos	2001-2008
15	Falências	Comparar modelos de previsão de falências com base na *accuracy* e o nº de regras	NE	NE
16	Falências	Conseguir novos modelos para classificar empresas iranianas cotadas falidas e não falidas utilizando *GEP* e *MEP*	7 anos	1999-2006
17	DF	Criar um modelo preditivo das DF com diversas técnicas (tradicionais e de *ML*), utilizando rácios financeiros e não financeiros	7 anos	1999-2006
18	DF	Criar um modelo a partir de um algoritmo para a previsão de DF das empresas cotadas	6 anos	2000-2005
19	Falências	Propor uma abordagem de *MCLP* para a prospeção de dados para a previsão de falências	7 anos	1992-1998
20	Falências	Encontrar o modelo de previsão do início da falência de uma empresa com o menor erro ao quadrado, comparando-os	1 ano	1996-1997
21	Falências	Estudar rácios financeiros como preditores da falência de empresas japonesas	NE	NE

Fonte: Elaboração própria

Notas: DF- Dificuldades Financeiras; *EWM- Early Warning Models; GEP- Gene Expression Programming; LIME- Local interpretable model-agnostic explanations; LSTM- Long short-term Memory*; *MCLP- Multiple Criteria Linear Programming; MEP- Multi-expression Programming; ML- Machine learning;* NE- Não Especificado; PM- Pequenas e Médias; PME- Pequenas e Médias Empresas; *RNN- Recurrent Neural Networks*; *ST- Special Treatment*

Quanto ao tipo de empresas estudados (Tabela 2.5), verifica-se que mais de metade dos artigos selecionados, 11 artigos de um total de 21, aplicam estudos em empresas cotadas pela importância que têm para os investidores (e.g., Sun e Li, 2008) e geograficamente estes estudos têm maior impacto na China, com cinco dos 21 artigos selecionados (e.g., Geng *et al.*, 2015; Zhou *et al.*, 2015).

Quanto à dimensão da empresa verificamos que apenas quatro dos 21 artigos aplicam o seu estudo a Pequenas e Médias Empresas (PME) (e.g., Bragoli *et al.*, 2022; Park *et al.*, 2021). No que se refere ao setor em estudo, a maioria não o identifica (e.g., Kwak *et al.*, 2004; Shirata e Terano, 2000). No entanto, três referem que aplicaram os seus estudos ao setor bancário/crédito (Foster e Stine, 2004; Malakauskas e Lakstutiene, 2021; Petropoulos *et al.*, 2020).

Por último, fazendo uma análise quantitativa dos dados, constata-se que mais de metade dos artigos selecionados, cerca de 66,7%, utiliza dados que não estão equilibrados (i.e., em que o número de empresas "saudáveis" é diferente do número de empresas que se encontram em dificuldades financeiras, insolvência ou falência) o que poderá causar dificuldades na avaliação dos modelos (e.g., Bragoli *et al.*, 2022; Kim *et al.*, 2022). Por fim, a base de dados mais utilizada para a recolha de dados foi a *China Stock Market and Accounting Research (CSMAR)*, utilizada em todos os artigos em que foram recolhidos dados da China, ou seja, em cinco dos selecionados (e.g., Sun e Li, 2008).

Tabela 2.5: Contextualização dos artigos da RSL- Entidades estudadas

ID	Tipo de empresas (PME?)	Empresas cotadas?	Setor	País da amostra	Base de dados	Nº entidades (balanceado?)
1	NE	NE	Indústria não-financeira	NE	*Compustat North America dataset; CRSP dataset*	454 752 (Não)
2	Não	C	5 (Comércio; Conglomerados; Indústrias; Propriedades; Utilidade Pública)	China	*CSMAR*	29 000 (2 métodos)
3	Sim	C	7 (Empresas manufatureiras- comida, têxtil, peles, madeira, vidro, metal e maquinaria)	Itália	*AIDA Bureau van Dijk*	70 309 (Não)
4	NE	NE	NE	Itália	*ISTAT*	7 795 (Não)
5	Sim	NE	NE	Coreia	*Douzone Bizon ICT Group*	546 619 (Não)
6	Sim	NE	1- Setor bancário	Estónia, Letónia e Lituânia	*Swedbank AB*	12 000 (Não)
7	Sim	NE	7 (Empresas manufatureiras- comida, têxtil, peles, madeira, vidro, metal e maquinaria)	NE	*KOSME*	4 358 (Não)
8	Não	C	NE	China	*CSMAR*	424 (Sim)
9	Não	C	1- Setor bancário	EUA	*FDIC*	11 573 (Não)
10	NE	NE	NE	NE	*University of California Irvine Machine Learning Database Repository*	250 (Não)
11	Não	C	16 (sendo os maiores materiais optoeletrónicos, semicondutores e componentes eletrónicos)	Tailândia	*TEJ*	64 (Sim)
12	Não	C	10 (Agricultura, comércio grossista, entre outros não especificados)	EUA e Canadá	*Salomon Center database; Compustat*	13 300 (Não)
13	NE	NE	NE	China	*GTA – CSMAR*	10 365 (Não)
14	Não	C	NE	China	*CSMAR*	214 (Sim)
15	NE	NE	NE	NE	NE	NE
16	Não	C	NE	Irão	*Tehran Stock Exchange*	136 (Não)
17	Não	C	NE	Tailândia	*TSEC*	68 (Sim)
18	Não	C	NE	China	*CSMAR*	198 (Não)
19	Não	C	NE	EUA	*Research Insight*	NE
20	NE	NE	1 (Indústria do crédito)	EUA	*Wharton Financial Institutions Center*	244 094 (Não)
21	NE	NE	NE	Japão	*Teikoku Data Bank Cosmos1 Database*	986 (Não)

Fonte: Elaboração própria

Notas: C- Cotadas; *CRSP- Center for Research in Security Prices; CSMAR- China Stock market and Accounting Research;* EUA- Estados Unidos da América*; FDIC- Federal Deposit Insurance Corporation; GTA- GuoTaiAn; ISTAT- Istituto Nazionale di Statistica; KOSME- Korea SMEs and Startups Agency;* NE- Não Específicado; *TEJ- Taiwan Economic Journal; TSEC- Taiwan stock exchange corporation*

2.1.2.2. Modelos utilizados na previsão e sua validação

Os artigos selecionados descrevem os diferentes algoritmos utilizados, as variáveis utilizadas nos modelos e a forma de seleção das mesmas (Tabela 2.6) e ainda os métodos de avaliação utilizados, as métricas de avaliação, os algoritmos que apresentam melhor performance, os tipos de gráficos utilizados, bem como as variáveis mais importantes (Tabela 2.7).

Relativamente aos algoritmos utilizados nos estudos, verifica-se que há uma grande diversidade. Dos estudos que recorrem a algoritmos de regressão, 91,7% dos artigos utilizam o algoritmo da regressão linear (e.g., Bragoli *et al.*, 2022; Tang *et al.*, 2020), sendo este o algoritmo mais utilizado; e dos estudos que recorrem a algoritmos de classificação, nove dos artigos utilizam árvores de decisão (e.g., Foster e Stine, 2004; Shirata e Terano, 2000).

Em termos da variável dependente, em quatro dos 21 artigos é estudada a variável binária das falências (e.g., Barboza *et al.*, 2017; Divsalar *et al.*, 2012). Adicionalmente, surge como variável dependente a variável binária das insolvências, em 1 dos 21 artigos (Lee *et al.*, 2020). Relativamente às variáveis independentes, as mais usadas foram os indicadores financeiros[4], em 76,2% dos artigos selecionados (e.g., Divsalar *et al.*, 2012; Sun e Li, 2008).

Naturalmente, cada estudo teve em conta um número de variáveis independentes diferente e cada artigo selecionou as variáveis adequadas para cada modelo de sua forma. Quanto ao número de variáveis escolhidas, o mínimo de variáveis utilizadas é quatro (Divsalar *et al.*, 2012), sendo que o máximo são 255 variáveis (Foster e Stine, 2004). A seleção destas variáveis, em 16 dos selecionados, foi feita maioritariamente a partir da literatura (Barboza *et al.*, 2017; Zhou *et al.*, 2015).

[4] Os indicadores financeiros utilizados nos diferentes estudos basearam-se em Ativo corrente/Total ativo; Ativo corrente/Total passivo; Fluxo de caixa/Financiamentos obtidos; Fluxo de caixa/Total ativo; Variação nº empregados; Financiamentos obtidos/Capital próprio; Fundo de maneio/Total ativo; rácio de endividamento; rácio de liquidez; Margem de lucro; rentabilidade do ativo; rentabilidade do capital próprio; rentabilidade das vendas; autonomia financeira; solvabilidade; entre outros.

Tabela 2.6: Características dos modelos utilizados nos artigos da RSL

ID	Problema	Algoritmo	Variáveis dependentes	Nº variáveis independentes	Dimensões das variáveis independentes
1	C; R	*RNN; LSTM; LR; RF e SVM; Ensemble model*	Binária: 1- falência; 0- caso contrário	8	IF
2	C; R	*Logit; GLM; GAM; Dynamic Hazard; DT; RF*	Binária: 1- dificuldades financeiras; 0- caso contrário	16	Rácios contabilísticos, de mercado e de crescimento
3	C; R	*LR; ANN; RF; XGBoost*	Binária: 1- falência; 0- caso contrário	24	IF e variáveis industriais/regionais
4	C; R	*LASSO; RF; ANN; Gradient Boosted Machine*	Binária: 1- falência; 0- caso contrário	26	IF e demográficos
5	C	*XGBoost; LightGBM*	NE	25	IF
6	C; R	*LR; ANN; RF*	Binária: 1- dificuldades financeiras; 0- caso contrário	23	IF
7	C; R	*LR; DT; ANN* e *ensemble model*	Binária: 1- insolvência; 0- caso contrário	37	Capacidade de gestão, viabilidade do negócio e capacidade técnica
8	C; R	*LR; SVM; DT; RF; GBDT; Xgboost; SVM; DNN; RNN; LSTM*	NE	74	IF; indicadores de gestão e textuais
9	C; R	*LR; LDA; RF; SVM; ANN; CRF*	Binária: 1- *default*; 0- caso contrário	23	*CAMELS*
10	C	*BPNN; PNN; RBFNN; GRNN*	Binária: falências e não falências	6	*Industrial risk, management risk, financial flexibility, credibility, competitiveness, and operating risk*
11	C	*SVM; HACT; Hybrid GA-fuzzy clustering e XGBoost; DBN*	Binária: 1- dificuldades financeiras; 0- caso contrário	16	IF
12	C	*SVM; boosting; bagging; RF (CART); ANN; LR; DA*	Binária: 0- falência; 1- caso contrário	11	IF
13	C; R	*LR; kNN; DT (C4.5); Ripper; ANN; SVM*	Binária: 1- dificuldades financeiras; 0- caso contrário	169	IF e variáveis de mercado
14	C	*ANN; DT; SVM*	Binária: 1- deve receber o rótulo de *ST*; 0-caso contrário	31	IF
15	C	*ANN; DT (ID3, C4.5, C5 e CART); LR; SVM*	Binária: falências e não falências	19	IF, *data year* e *CIK number*
16	C	*GEP* e *MEP; LR; LSR*	Binária: 1- falência; 0- caso contrário	4	IF
17	C	*ANN; Clusters (K-means)*	Binária: falências e não falências	13	IF e não financeiros
18	C	*AOI; IG; DT*	Binária: em dificuldades financeiras e sem dificuldades financeiras	35	IF
19	R	*MCLP*	Binária: falências e não falências	5	IF
20	C; R	*DT (C4.5 and C5.0); LSR*	Binária: falências e não falências	255	NE
21	C; R	*DT (C 4.5, SIBILE, CART); LR; Stepdisc*	NE	66	IF e tipo de indústria

Fonte: Elaboração própria

Notas: *ANN- Artificial Neural Networks; AOI- Attribute-Oriented Induction; BPNN- Back Propagation Neural Network;* C- Classificação; *CAMELS- Capital, Asset Quality, Management, Earnings, Liquidity, and Sensitivity; CART- Classification and Regression Trees; CIK- Central Index Key; CRF- Random Forests of Conditional Inference Trees; DA- Discriminant Analysis; DBN- Deep Belief Network; DNN- Deep Neural Network; DT- Decision Tree; GA- Genetic Algorithm; GAM- Generalized Additive Model; GBDT- Gradient Boosting Decision Tree; GEP- Gene Expression Programming; GLM- Generalized Linear Model; GRNN- Generalized Regression Neural Network; HACT- Hybrid Associative Classifier with Translation;* IF- Indicadores Financeiros; *IG- Information Gain; kNN- k-nearest Neighbors; LASSO- Least Absolute Shrinkage and Selection Operator; LDA- Linear Discriminant Analysis; LR- Logistic Regression; LSR- Least squares regression; LSTM- Long short-term memory; MCLP- Multiple criteria linear programming; MEP- Multi-expression Programming; MLP- Multi-layer Perceptron;* NE- Não Especificado; *PNN-Probabilistic Neural Network;* R- Regressão; *RBF- Radial Basis Function; RBFNN-Radial Basis Neural Network; RF- Random Forest; RNN- Recurrent Neural Network; ST- Special Treatment; SVM- Support Vector Machine*

Na ótica da avaliação dos modelos estudados (Tabela 2.7), três da totalidade dos artigos utilizam o *10-fold cross validation* (validação cruzada com 10 partições de semelhante dimensão) para a avaliação dos modelos (Lahmiri e Bekiros, 2019; Zhou *et al.*, 2015) e utilizam diferentes métricas de avaliação, sendo que a mais comum nos artigos selecionados é a *accuracy*, utilizada em 16 dos artigos. Esta métrica tem os seus valores a variar entre 9,75% (Yousaf *et al.*, 2022) e 99,36% (Kim *et al.*, 2022). Tendo em conta as diferentes métricas de avaliação, em quatro artigos eleitos, o modelo com melhor performance foi o *Random Forest (RF*) (e.g., Antulov-Fantulin *et al.*, 2021; Malakauskas e Lakstutiene, 2021).

Quanto às variáveis mais importantes (i.e., variáveis que melhor ajudam a explicar a variável dependente de cada modelo), apenas 10 dos 21 artigos as identificam (Huang e Yen, 2019; Petropoulos *et al.*, 2020), sendo que em 70% dos artigos foram os indicadores financeiros os preditores mais importantes (Antulov-Fantulin *et al.*, 2021; Huang e Yen, 2019).

Tabela 2.7: Avaliação dos modelos utilizados nos artigos da RSL

ID	Método de avaliação	Métricas de avaliação	Melhor performance	Variáveis mais importantes
1	NE	A (72,36%-99,36%); P (0,01%-0,58%); S (21,74%-47,83%); F1-Score (0,0002-0,0114); AUC (0,4872-0,7305)	*Ensemble model*	NE
2	NE	A (9,75%-94,72%); S (100,00%); Sp (9,72%-94,72%); AUC (0,99)	*RF*	NE
3	NE	S (84,63%-89,73%)	*XGBoost*	NE
4	*Cross-validation*	P (0,01%-11,7%); AUC (0,500-0,983)	*RF*	10 (Indicadores financeiros e demográficos)
5	*5-fold cross validation*	AUC (0,876-0,919)	*Light-GBM*	NE
6	*Cross-validation and stratified sampling techniques*	A (52,01%-52,89%); S (60,69%-65,83%); Sp (51,11%-51,58%)	*RF*	3 *(*Indicadores financeiros e idade da empresa)
7	NE	A (<3 anos: 59,30%-69,10%; >3 anos: 62,80%-82,70%)	*Boosted DT*	3 (indicadores financeiros, fiabilidade do CEO e capacidade competitiva)
8	*10-fold cross validation*	A (73,81%-85,70%)	*DNN*	26 (Indicadores financeiros e de gestão)
9	*10-fold cross validation*	A (77,40%-99,20%)	*RF*	16 (Indicadores de rentabilidade e de capital)
10	*10-fold cross-validation*	A (84,81%-99,96%); S (51,78%-100,00%); Sp (58,94%-100,00%)	*GRNN*	NE
11	*4-fold cross validation*	A (70,30%-90,60%); S (85,90%-93,70%)	*XGBoost*	6 (Indicadores financeiros)
12	NE	A (52,18%-87,06%); S (52,05%-86,71%)	*Boosting, Bagging and RF*	NE
13	*10-fold cross-validation*	A (79,64%-93,97%); S (70,62%-98,60%); Sp (67,38%-98,22%)	*LR*	NE
14	*Cross-validation*	A (71,10%-78,80%); S (72,80%-79,90%); P (63,60%-73,20%)	*ANN*	10 *(*Indicadores financeiros e demográficos*)*
15	*10-fold cross validation*	A (60,90%-94,80%)	DT	NE
16	NE	A (79,41%-91,18%); S (80,00%-95,00%); Sp (78,57%-85,71%)	*GEP* e *MEP*	5 (Indicadores financeiros)
17	NE	A (60,00%-82,14%)	*ANN*	NE
18	*Method of resubstitution, 10-fold cross-validation*	A (81,25%-95,33%)	DT	NE
19	NE	S (71,00%-100,00%); Sp (81,00%-100,00%)	*MCLP*	NE
20	*5-fold cross-validation*	N.E.	*LSR*	NE
21	NE	A (85,00%-87,00%)	*LR*	10-19 (indicadores financeiros)

Fonte: Elaboração própria

Notas: *ANN- Artificial Neural Network; AUC- Area under the ROC (Receiver operating characteristic) curve*; *DNN- Deep Neural Network; DT- Decision Tree*; *GEP- Gene Expression Programming; GRNN- Generalized Regression Neural Network; LR- Logistic Regression; LSR- Least squares regression; MCLP- Multiple criteria linear programming; MEP- Multi-expression Programming;* NE- Não Específicado; *RF- Random Forest; WBA- Weighted Balance accuracy*

2.1.2.3. Avaliação dos resultados dos estudos

Com a intenção de avaliar os resultados obtidos nos estudos, a Tabela 2.8 sistematiza a totalidade dos 21 artigos da RSL. São alvo de análise as limitações dos estudos, os seus contributos, as pistas de investigação futura e o *stakeholder* destacado no estudo.

Quanto aos contributos (Tabela 2.8), estes são variados, sendo que grande parte dos artigos contribuiu para a literatura com algoritmos *machine learning* (e.g., Kim *et al.*, 2022; Lee *et al.*, 2020) ou expandiram os fatores associados a falências, insolvências e dificuldades financeiras, para além de indicadores financeiros (e.g., Antulov-Fantulin *et al.*, 2021; Tang *et al.*, 2020).

Quanto às limitações, muitos dos artigos não as especificaram. No entanto, nos que o fizeram, o problema centrou-se: nos dados não equilibrados (Antulov-Fantulin *et al.*, 2021); variáveis presentes na literatura que não foram selecionadas pelos métodos de escolha de variáveis utilizados (e.g., Tang *et al.*, 2020; Yousaf *et al.*, 2022); dificuldades em perceber o *output* do modelo quanto ao intervalo de tempo a que se referia (Kim *et al.*, 2022); dificuldade na recolha de dados de PME e pequena amostra da classe minoritária (Lee *et al.*, 2020); e muita variabilidade nas métricas de avaliação dos resultados utilizando algoritmos de *ANN* (Barboza *et al.*, 2017).

Quanto às pistas de investigação futura (Tabela 2.8), estas surgiram em grande parte como consequência das limitações dos diferentes estudos. Há autores que sugerem amplificar os modelos para vários períodos, e não apenas para a previsão do período seguinte (e.g., Kim *et al.*, 2022), outros recomendam a inclusão de indicadores não financeiros (Antulov-Fantulin *et al.*, 2021; Geng *et al.*, 2015).

Por último, quanto aos *stakeholders* destacados em cada artigo devido aos seus contributos para com as empresas, em 12 dos artigos, são aos investidores (e.g., Kim *et al.*, 2022; Sun e Li, 2008) que os estudos demonstram uma maior atenção; seguidos dos credores, em cerca de 10 dos artigos selecionados (e.g., Antulov-Fantulin *et al.*, 2021; Park *et al.*, 2021).

Tabela 2.8: Avaliação dos resultados dos estudos

ID	Contributos	Investigações futuras	*Stakeholder* destacado
1	Utilização de técnicas de *ML*	Contemplar modelos multiperíodos	Investidores
2	Maior período; Consideração de variáveis de crescimento	Incluir indicadores não financeiros	Investidores
3	NE	NE	Credores, investidores, entidades reguladoras, *PM* e gestores
4	Características não financeiras mais importantes do que financeiras	Incluir mais indicadores não financeiros	Credores e investidores
5	Aplicação do método *LIME* a modelos de *black box* como o *XGBoost* e *LightGBM*	Combinar o modelo utilizado com *NNRW* para construir um modelo capaz de ser extrapolado	Credores, investidores e bancos
6	Melhoria de tomada de decisão dos financiadores	NE	*Decision makers*
7	Uso de técnicas de *data mining*	Avaliar a relação entre fundos governamentais e insolvências das PME no ambiente da COVID-19	Acionistas, investidores e credores
8	Novos preditores de dificuldades financeiras	Avaliar os efeitos do aumento do nº de empresas saudáveis no treino; uso de um período mais longo	Acionistas, empresas cotadas e a própria economia
9	Utilização de 6 algoritmos; métricas de *performance* para dados desequilibrados; mais variáveis independentes	Construção de um sistema global de *rating* para os bancos	Investidores e reguladores
10	Comparação entre 4 tipos de *ANN* quando se consideram dados qualitativos	Otimizar os dados com uma técnica particular de otimização heurística	Acionistas, credores, *SM*, auditores, e AP
11	Nova perspetiva do desempenho relativo da previsão de dificuldades financeiras	Melhoramento das técnicas de *ML* a partir da função de extração de características de um *DBN*	Credores e acionistas
12	Variedade de técnicas e a aplicabilidade do modelo	Incorporação de taxas de crescimento e/ou efeitos temporais; Aplicação em instituições financeiras	Investidores e *lender institutions*
13	Utilização de mais de 300 variáveis; e abordagem de *data mining*	Papel da informação atualizada de uma empresa na precisão da previsão	Investidores, credores e *partners*
14	Utilização de novas variáveis (e.g., margem de lucro líquido, retorno do ativo total, ganhos por ação)	Utilização de indicadores não financeiros no modelo de previsão; uso de diferentes países e comparação e contraste entre estes modelos	Investidores e gestores
15	Controlo do *tradeoff* entre *accuracy* média e profundidade da *DT*	NE	Acionistas, credores, *PM* e gestores
16	Identificação dos rácios financeiros com base na literatura e numa análise sequencial de seleção de características	NE	NE
17	Comparação de método estatístico tradicional com a abordagem de inteligência artificial	Aplicação de técnicas adicionais de inteligência artificial; Expansão do sistema de modo a lidar com mais conjuntos de dados financeiros	Agentes de empréstimos bancários, credores acionistas, obrigacionistas, analistas financeiros, funcionários governamentais e público em geral
18	Aumento da capacidade dinâmica de aprendizagem das DF de empresas cotadas	Adotar o modelo em empresas cotadas	Investidores
19	Modelo com melhor performance que os tradicionais	Comparar a abordagem de *data mining MCLP* com a *DT* para ver que abordagem é mais eficiente e eficaz na previsão de falências	Investidores, credores, auditores, gestores, sindicatos, empregados e público em geral
20	NE	NE	NE
21	Seleção de características com inteligência artificial melhor do que com métodos estatísticos	NE	NE

Fonte: Elaboração própria

Notas: *ANN- Artificial Neural Networks;* AP- Autoridades Públicas*; DBN- Deep Belief Network; DT- Decision tree; LIME- Local Interpretable Model-Agnostic Explanations; MCLP- Multiple Criteria Linear Programming; ML- Machine Learning; NNRW-Neural Network with Random Weights;* PME- Pequenas e Médias Empresas*; PM- Policy makers*; *SM- Senior Managers*

2.1.2.4. Avaliação da qualidade dos artigos

Tendo este estudo por base uma RSL, é importante facilitar o trabalho a investigadores e profissionais. Assim, avalia-se a relevância dos artigos em cada uma das três dimensões apresentadas no protocolo da RSL (Tabela 2.3). Para cada critério de qualidade é atribuído uma pontuação a cada pergunta de avaliação e para cada artigo: 0 se não responde à questão; 0,5 se responde parcialmente e 1 se responde totalmente. A Tabela 2.9 apresenta as pontuações dos 21 artigos e as cotações para cada item de avaliação e dimensão.

Dos artigos incluídos nesta RSL, conclui-se que nenhum artigo tem pontuação máxima (13 pontos) e que apenas três dos artigos analisados obtém maior pontuação em termos de qualidade, 10 pontos (Barboza *et al.*, 2017; Petropoulos *et al.*, 2020; Yousaf *et al.*, 2022), sendo que o critério de avaliação que apresenta maior qualidade é o "Q2.2: Apresenta, descreve e justifica a(s) técnica(s)/ algoritmo(s) utilizada/o (s)?".

A dimensão que se encontra com melhor avaliação é a dos modelos utilizados na previsão das falências/insolvências/dificuldades financeiras e sua validação, tendo conseguido a maior média de pontuação 13,8, para um máximo de 21. Aliás, é nesta dimensão onde se encontram os dois critérios com melhor pontuação (Q2.2 e Q2.3), em que são alvo de análise os algoritmos utilizados e a comparação entre eles. Em oposição, surge a primeira dimensão a ser avaliada, âmbitos e contextos da realização dos estudos, onde a diferença entre insolvências, falências e dificuldades financeiras (Q1.2) surge com uma avaliação média muito baixa (1,5, em que apenas dois artigos abordam estes aspetos (Geng *et al.*, 2015; Malakauskas e Lakstutiene, 2021) Neste sentido, conclui-se que poucos são os artigos que definem claramente insolvências, falências e dificuldades financeiras (Q1.1) e que utilizam técnicas de robustez e de sensibilidade (Q2.5).

Em suma, para a dimensão dos âmbitos e contextos da realização dos estudos, recomenda-se a leitura do artigo de Malakauskas e Lakstutiene (2021), para a dimensão dos modelos utilizados na previsão das falências/insolvências/dificuldades financeiras e sua validação os artigos, em geral, Petropoulos *et al.* (2020); Tang *et al.* (2020); Yousaf *et al.* (2022) e, por fim, para a dimensão da avaliação dos resultados os artigos, em geral, de Barboza *et al.* (2017); Lee *et al.* (2020); Petropoulos *et al.* (2020).

Tabela 2.9: Avaliação da qualidade dos artigos da RSL

ID	Âmbitos e contextos da realização dos estudos					Modelos utilizados na previsão das falências/insolvências/dificuldades financeiras e sua validação					Avaliação dos resultados			Total
	Q1.1	Q1.2	Q1.3	Q1.4	Q1.5	Q2.1	Q2.2	Q2.3	Q2.4	Q2.5	Q3.1	Q3.2	Q3.3	
1	0,5	0	1	0,5	0,5	1	1	1	0,5	0	1	1	1	**9**
2	0	0	1	1	1	1	1	1	1	0	1	1	1	**10**
3	0	0	1	0,5	0,5	0,5	1	1	0,5	0	0,5	0	0	**5,5**
4	0	0	1	0,5	0,5	1	1	1	0,5	0	0,5	0,5	0	**6,5**
5	0	0	0,5	0,5	0,5	0,5	1	1	0,5	0	1	1	0	**6,5**
6	1	1	1	1	0,5	1	1	1	0,5	0	0,5	0,5	0	**9**
7	0,5	0	1	0,5	0,5	0,5	1	1	0,5	0	1	1	1	**8,5**
8	0	0	1	1	1	1	1	1	1	0	0,5	1	1	**9,5**
9	0	0	1	1	1	1	1	1	0,5	0,5	1	1	1	**10**
10	0	0	0,5	0	0,5	1	1	1	0,5	0,5	0,5	1	0	**6,5**
11	1	0	1	1	0,5	0,5	1	1	1	0	1	1	0	**9**
12	0	0	1	1	1	1	1	1	0,5	0,5	1	1	1	**10**
13	0	0	1	0,5	0,5	0,5	1	1	0,5	0	0,5	1	0	**6,5**
14	1	0,5	1	0,5	0,5	1	1	1	0,5	0,5	1	1	0	**9,5**
15	0	0	0,5	0	0,5	0,5	0,5	0,5	0,5	0	0,5	0,5	0	**4**
16	1	0	1	0,5	0,5	1	1	1	0,5	0	1	1	0	**8,5**
17	1	0	1	0,5	0,5	1	1	1	0,5	0	1	1	0	**8,5**
18	0	0	1	1	0,5	0,5	0,5	0	0,5	0	0	0,5	0	**4,5**
19	0	0	1	0,5	0,5	1	1	1	0,5	0	0,5	0	0	**6**
20	0	0	0,5	0,5	0,5	0,5	0,5	0,5	0	0	0,5	0	0	**3,5**
21	0	0	0,5	0,5	0,5	1	1	1	0,5	0	0,5	0	0	**5,5**
Total	**6**	**1,5**	**18,5**	**13**	**12,5**	**17**	**19,5**	**19**	**11,5**	**2**	**15**	**15**	**6**	**156,5**

Fonte: Elaboração própria

2.2. Contributos e limitações da revisão

A revisão de literatura pretende aumentar o grau de conhecimento do tema de predição das insolvências, falências e dificuldades financeiras através da utilização de técnicas avançadas de análise de dados (*advanced analytics*) e utilizando indicadores financeiros e não financeiros. O tema das insolvências ainda se trata de um *gap* na literatura uma vez que ainda existem poucos estudos publicados em *journals* indexados no *WoS* sobre o mesmo, tal como sugere Cathcart *et al.* (2020). Como tal, é um bom tema para investigar e desenvolver, especialmente se forem considerados também indicadores não financeiros.

É importante ressalvar que nos estudos realizados até ao momento já foram desenvolvidos modelos com elevada capacidade preditiva (e.g., Kim *et al.*, 2022; Lahmiri e Bekiros, 2019). Não obstante ainda existe margem para melhorias, incluindo indicadores não financeiros como a opinião do auditor (quando aplicável) (Stanisic *et al.*, 2013); e o género da direção, já explorado por Wilson *et al.* (2014). Relativamente às variáveis não financeiras, as mesmas são pouco utilizadas na predição das insolvências, apesar da sua elevada capacidade para distinguir empresas ativas e insolventes, como evidenciado no estudo de Antulov-Fantulin *et al.* (2021).

Os modelos preditivos baseados em árvores de decisão (ou regras) permitem identificar facilmente as causas (variáveis) da situação débil em que se encontra a empresa e, assim, identificar e implementar medidas que possam reverter as suas preocupações. Por outro lado, no que diz respeito aos utilizadores da informação, nomeadamente os investidores, credores e bancos, um modelo preditivo permite-lhes identificar as empresas com propensões em entrar em insolvência de forma a tomarem melhores decisões de investimento e financiamento (Park *et al.*, 2021).

Importa evidenciar que alguns estudos tiveram em conta um número elevado de variáveis independentes (e.g., Foster e Stine, 2004) o que torna a investigação demasiado complexa; e que alguns dos estudos não tomam em consideração indicadores não financeiros (e.g., Kim *et al.*, 2022; Malakauskas e Lakstutiene, 2021).

No que diz respeito a investigações futuras e com o intuito de colmatar as limitações evidenciadas anteriormente, sugere-se a adoção de outras variáveis não financeiras tal como os indicadores mencionados anteriormente, com o intuito de perceber se a inclusão destas variáveis aumenta a capacidade preditiva do modelo bem como a utilização de um número modesto de variáveis independentes. No capítulo seguinte será abordada em maior detalhe a metodologia adotada.

3. Metodologia *Cross Industry Standard Process for Data Mining*

Este capítulo apresenta o posicionamento da investigação e explicita a metodologia adotada, *Cross Industry Standard Process for Data Mining* (*CRISP-DM*).

Esta investigação baseia-se numa pesquisa positivista, que tomou uma posição dominante a partir da década de 1970, e permite observar a relação entre diferentes variáveis explicativas com a construção de modelos estatísticos para testar estas relações (Major, 2017). Assim, de acordo com os objetivos estipulados, este estudo segue uma abordagem mista utilizando variáveis quantitativas e qualitativas.

Noutra perspetiva, este estudo é considerado um estudo exploratório, uma vez que aborda o tema da previsão de insolvências e falências utilizando variáveis financeiras e não financeiras como preditores. Ao fazê-lo, contribui para a obtenção de conhecimento sobre este tema pouco abordado na literatura.

O *CRISP-DM* é um processo que conduz a metodologia de *data mining*. De acordo com Lebkiri *et al.* (2021) este tipo de processo é o mais usado entre estudos do género e, como tal, sem exceção, é levado a cabo nesta investigação. O *CRISP-DM* envolve seis fases distintas: (1) compreensão do negócio; (2) compreensão dos dados; (3) preparação dos dados; (4) modelação; (5) avaliação; e (6) implementação.

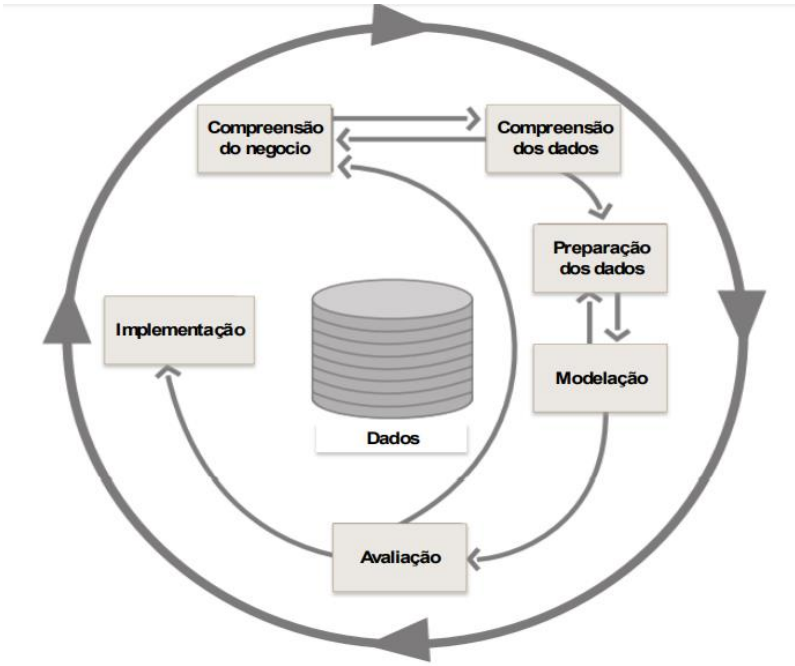

Figura 3.1: Processo do *CRISP-DM*

Fonte: Peixoto (2015, p. 18)

A primeira fase é de grande relevância visto que trata a identificação dos objetivos da investigação e a elaboração do plano de trabalho para chegar aos objetivos pretendidos. Posteriormente na segunda fase, extraem-se os dados, avalia-se a sua qualidade e a ligação ao objetivo proposto. De seguida, na terceira fase, designada preparação dos dados, são feitas todas as tarefas necessárias de preparação e limpeza dos dados para serem utilizados na fase final da amostra. Na quarta e penúltima fase, denominada modelação, são aplicados diferentes algoritmos de *machine learning* aos dados recolhidos. Por fim, na quinta fase, designada avaliação, a qualidade do modelo obtido é avaliada e o alcance dos objetivos inicialmente propostos analisado. Por fim, a sexta fase, denominada implementação, engloba a implementação prática do modelo, assim como o controlo do mesmo. Esta tarefa encontra-se fora do âmbito desta dissertação, embora este documento seja uma forma de implementação, ao descrever detalhadamente a forma de obtenção dos modelos e suas interpretações.

3.1. Compreensão do negócio

Como visto anteriormente, o objetivo geral desta investigação é identificar o impacto que as estruturas financeira e societária de uma empresa têm na explicação das insolvências e das falências. Para alcançar esse objetivo, são estabelecidos três objetivos específicos que englobam a execução de diferentes tarefas, visando atingir os objetivos propostos.

Para a concretização do primeiro objetivo, é feita uma análise bivariada das variáveis independentes tendo em conta o *status* das empresas (ativa ou não ativa). Com o objetivo de se encontrar diferenças significativas entre os dois tipos de empresas, utilizando o nível de significância mais usual (i.e., 5%) de acordo com Laureano (2020).

Para a realização do segundo objetivo, são elaborados diversos modelos preditivos do *status* utilizando, de acordo com a revisão da literatura, a técnica de árvores de decisão devido à sua elevada capacidade de perceção por parte de quem a interpreta (Olson *et al.*, 2012). Considera-se que se tem um bom modelo preditivo quando as suas métricas de avaliação são superiores às obtidas na RL (Tabela 3.1) visto que os autores consideram os seus modelos preditivos como bons modelos.

Tabela 3.1: Métricas de avaliação obtidas na revisão de literatura

Métricas	Valores obtidos
Accuracy	77,64%
Precisão	24,85%
Sensibilidade	77,66%
Especificidade	73,08%
F1-Score	0,01
AUC	0,81

Fonte: Elaboração própria

Por fim, para se identificar os perfis de empresas propensas à insolvência/falência e ativas, é realizada uma análise dos nós finais da árvore de decisão que apresentam os resultados com maior confiança (acima de 80%) e o suporte mais relevante (mais de 1.000 empresas).

Relativamente às ferramentas de análise de dados, atendendo à fácil aprendizagem e utilização, foram utilizadas: para a compreensão e preparação de dados, o Microsoft Excel; e para a modelação, o IBM SPSS Statistics (versão 27) e IBM SPSS Modeler (versão 18).

3.2. Compreensão e preparação dos dados

Os dados utilizados na presente investigação são extraídos da *Orbis* Europa (https://login.bvdinfo.com/R0/Orbis4Europe) devido à facilidade de acesso. Para obter a amostra inicial são definidos três critérios de pesquisa: (1) somente sociedades por quotas de responsabilidade limitada (Lda.) ou sociedades anónimas; (2) localizadas em Portugal; e (3) empresas ativas, insolventes e falidas. Desta forma, foi extraída uma amostra de 772.142 empresas juntamente com a respetiva informação financeira e societária.

Após a extração desta amostra é efetuado o tratamento dos dados (i.e., junção dos vários ficheiros obtidos), cálculo de rácios financeiros, proporção dos dados e tratamento de empresas com falta de dados e com existência de *outliers* em alguns dados financeiros. Devido à falta de dados relativamente a empresas em dificuldades financeiras, não foi possível extrair informações acerca destas empresas. De maneira a diminuir a dimensão dos dados, são consideradas dez vezes mais empresas ativas do que não ativas, tendo em conta o último ano de contas disponíveis. Os *outliers* são valores fora do padrão da amostra e, como tal, causam distorção nos resultados estatísticos obtidos (Dash *et al.*, 2023), pelo que são corrigidos tendo em conta um método denominado de winsorização, um processo que reduz os efeitos destes na amostra a analisar. Com este método, no conjunto de todas as variáveis incluídas, qualquer

valor acima do percentil 99 ou abaixo do percentil 1 é substituído pelo valor do referido percentil, respetivamente (Dash *et al.*, 2023), diminuindo assim o impacto dos *outliers* nos modelos preditivos. É de referir que se selecionaram os percentis 1 e o 99 por forma a não alterar muito os valores recolhidos (Yousaf *et al.*, 2022). Quanto às empresas para as quais todo o intervalo de tempo recolhido não tem dados acerca das variáveis, as mesmas são eliminadas da base de dados, não afetando a representatividade da amostra e mantendo 707.291 empresas.

A amostra selecionada é composta por 707.291 empresas com o último ano de contas disponíveis entre 2013 e 2022. Primeiramente, a Tabela 3.2 apresenta a distribuição das empresas pertencentes à amostra em termos de setor de atividade. Esta distribuição tem como base a Nomenclatura Estatística das Atividades Económicas da Comunidade Europeia (NACE). O setor com maior manifestação neste estudo é o setor comércio por grosso e a retalho; reparação de veículos automóveis e motociclos, representando cerca de 24,90% das empresas, seguindo-se do setor das indústrias transformadoras (10,70%).

Tabela 3.2: Distribuição das empresas por setor de atividade

Setor	Descrição	Nº	%
1	Agricultura, floresta e pesca	27.439	3,90%
2	Indústrias extrativas	1.326	0,20%
3	Indústrias transformadoras	75.402	10,70%
4	Produção e distribuição de eletricidade, gás, vapor e ar frio	1.414	0,20%
5	Captação, tratamento e distribuição de água; saneamento, gestão de resíduos e despoluição	1.747	0,20%
6	Construção	74.522	10,50%
7	Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos	175.858	24,90%
8	Transporte e armazenagem	36.169	5,10%
9	Atividades de alojamento e restauração	57.123	8,10%
10	Informação e comunicação	19.124	2,70%
11	Atividades financeiras e de seguros	12.241	1,70%
12	Atividades imobiliárias	50.620	7,20%
13	Atividades de consultoria, científicas, técnicas e similares	74.459	10,50%
14	Atividades administrativas e dos serviços de apoio	23.398	3,30%
15	Administração pública e defesa; segurança social obrigatória	20	0,00%
16	Educação	8.815	1,20%
17	Saúde humana e ação social	46.808	6,60%
18	Atividades artísticas, de espetáculos e recreativas	9.937	1,40%
19	Outras atividades de serviços	10.869	1,50%
	Total	**707.291**	**100,00%**

Fonte: Elaboração própria

Por outro lado, a Tabela 3.3 evidencia quais as regiões com mais empresas dentro da amostra selecionada. Predominantemente, as empresas localizam-se no Norte (n = 248.187; 35,09%) e na Área Metropolitana de Lisboa (AML) (n = 210.954; 29,83%) que contrasta com as empresas sediadas no Arquipélago dos Açores (n = 9.609; 1,36%). A tabela apresenta também a distribuição por dimensão, utilizando a classificação da base de dados Orbis[5],

[5] A partir da Orbis, para se definir como empresa muito grande, trata-se de uma empresa com um resultado operacional superior a 100.000.000€, um ativo superior a 200.000.000€, número de empregados superior a 1.000 e que seja uma empresa cotada. Uma empresa grande, trata-se de uma empresa com um resultado operacional superior a 10.000.000€, um ativo superior a 20.000.000€, número de empregados superior a 150 e que não seja classificada como muito grande. Uma empresa média, trata-se de uma empresa com um resultado operacional superior a 1.000.000€, um ativo superior a 2.000.000€, número de empregados superior a 15 e que não seja classificada como muito grande ou grande. Já uma pequena empresa é aquela que não tenha sido classificada como muito grande, grande ou média.

constatando-se que as pequenas empresas assumem uma maior preponderância (n = 568.097; 80,32%), sendo que as grandes empresas representam apenas (n = 15.580; 0,29%) da amostra, o que seria de esperar visto que 99,30% das empresas em Portugal são micros e pequenas empresas (Pordata, 2023).

Tabela 3.3: Distribuição das empresas por região e dimensão

Região/Dimensão	Pequena		Média		Grande		Muito grande		Total	
	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%
Norte	196.860	34,65%	45.705	37,61%	5.050	32,41%	572	27,43%	248.187	35,09%
Centro	118.780	20,91%	24.616	20,26%	3.183	20,43%	291	13,96%	146.870	20,77%
AML	170.937	30,09%	33.596	27,64%	5.387	34,58%	1.034	49,59%	210.954	29,83%
Alentejo	35.060	6,17%	7.248	5,96%	886	5,69%	67	3,21%	43.261	6,12%
Algarve	26.796	4,72%	5.275	4,34%	387	2,48%	24	1,15%	32.482	4,59%
Madeira	12.328	2,17%	3.035	2,50%	490	3,15%	75	3,60%	15.928	2,25%
Açores	7.336	1,29%	2.054	1,69%	197	1,26%	22	1,06%	9.609	1,36%
Total	**568.097**	**100,00%**	**121.529**	**100,00%**	**15.580**	**100,00%**	**2.085**	**100,00%**	**707.291**	**100,00%**

Nota: AML: Área Metropolitana de Lisboa

Fonte: Elaboração própria

3.2.1. Variável dependente

A variável dependente, que se pretende explicar e prever, é uma variável binária que diferencia empresas com *status* "Ativa" (n = 642.983; 90,91%) das empresas com *status* "Não ativa" (n = 64.308; 9,09%). Este tipo de variável é utilizado em grande parte dos estudos analisados na RL, em que o objetivo dos mesmos é a construção de modelos preditivos de insolvências/falências (Lee *et al.*, 2020). A Tabela 3.4 evidencia que existem 10 vezes mais empresas ativas do que não ativas e apresenta a sua distribuição pelo último ano de contas disponíveis.

Tabela 3.4: Distribuição das empresas por último ano de contas disponíveis

		Ativa	Não ativa	Total
2013	**Nº**	77.366	7.738	85.104
	%	90,91%	9,09%	100,00%
2014	**Nº**	85.671	8.568	94.239
	%	90,91%	9,09%	100,00%
2015	**Nº**	81.332	8.134	89.466
	%	90,91%	9,09%	100,00%
2016	**Nº**	71.652	7.166	78.818
	%	90,91%	9,09%	100,00%
2017	**Nº**	66.704	6.671	73.375
	%	90,91%	9,09%	100,00%
2018	**Nº**	69.049	6.906	75.955
	%	90,91%	9,09%	100,00%
2019	**Nº**	59.016	5.902	64.918
	%	90,91%	9,09%	100,00%
2020	**Nº**	57.146	5.715	62.861
	%	90,91%	9,09%	100,00%
2021	**Nº**	75.047	7.505	82.552
	%	90,91%	9,09%	100,00%
2022	**Nº**	0	3	3
	%	0,00%	100,00%	100,00%
Total	**Nº**	**642.983**	**64.308**	**707.291**
	%	**90,91%**	**9,09%**	**100,00%**

Fonte: Elaboração própria

3.2.2. Variáveis independentes

Quanto às variáveis independentes utilizadas nesta investigação, estas dividem-se em sete tipologias: demográficas, rentabilidade, crescimento, endividamento, estrutura do ativo, liquidez e rotação. Esta classificação tem em conta o estudo efetuado no artigo de Delen *et al.* (2013) e as variáveis são selecionadas tendo por base a RL bem como a facilidade/possibilidade de acesso e obtenção das mesmas.

À exceção das variáveis demográficas que são descritas tendo em conta o último ano de contas, os restantes seis grupos de variáveis são descritos tendo em conta o ano anterior ao

último ano de contas e são evidenciadas as suas fórmulas de cálculo[6]. Isto porque no caso das empresas não ativas, o último ano de contas por vezes já tem dados em falta e já não é viável a sua análise visto que já se previa a cessão de atividades nesse ano (i.e., último ano de contas). Quanto aos anos anteriores, o cálculo das medidas descritivas foi realizado, mas devido à extensão desta análise, a mesma não foi incluída nesta dissertação. Contudo, não se revela problemático uma vez que não se identificam variações significativas ao longo dos anos.

Em termos de variáveis demográficas, são englobadas sete. Para além do setor de atividade, região de localização e dimensão, já caracterizadas na Tabela 3.2 e Tabela 3.3 destacam-se:

- Anos_atividade: reflete a idade da empresa (em anos);
- Unipessoal: toma o valor de 1 se a empresa é unipessoal (i.e., é detida apenas por um sócio) e o valor de 0 se a empresa é coletiva (i.e., empresa tem mais do que um sócio);
- FormaJuridica: toma o valor de 1 se a sociedade é por quotas de responsabilidade limitada (Lda) e o valor de 0 se é uma sociedade anónima[7];
- Proporcao_Feminina: representa a proporção das mulheres no total de indivíduos presentes no conselho de administração em cada empresa, com o objetivo de haver um entendimento da relação entre o género feminino e o total dos membros do conselho de administração.

Em termos de variáveis de rentabilidade, são englobadas seis:

- ROAtivo (rentabilidade do ativo tendo em conta o resultado operacional): mede o resultado operacional gerado por cada unidade de ativo;
- RLAtivo (rentabilidade do ativo): mede a capacidade do ativo de gerar lucro. Este rácio difere do anterior visto que tem em conta gastos de depreciação/amortização,

[6] Ver no Anexo A – Medidas descritivas dos indicadores (página 61) e no Anexo B – Fórmulas de cálculo das variáveis independentes (página 64).

[7] Uma sociedade por quotas coletiva, segundo o artigo 197º do Código das Sociedades Comerciais (CSC), é uma sociedade cujo "capital está dividido em quotas e os sócios são solidariamente responsáveis por todas as entradas convencionadas no contrato social". O capital mínimo deste tipo de sociedades, de acordo com o artigo 201º do CSC "é livremente fixado no contrato de sociedade, correspondendo à soma das quotas subscritas pelos sócios". Já uma sociedade por quotas unipessoal, segundo o artigo 270º-A do CSC, é uma sociedade "constituída por um sócio único, pessoa singular ou colectiva, que é o titular da totalidade do capital social". As sociedades anónimas, de acordo com o artigo 271º do CSC, "o capital é dividido em acções e cada sócio limita a sua responsabilidade ao valor das acções que subscreveu.". De acordo com o artigo 273º do CSC, o número de sócios não pode ser inferior a 5, e além disso, tendo em conta o artigo 276º do CSC, o valor mínimo do capital é de 50.000€.

imparidade de ativos depreciáveis/amortizáveis, juros obtidos e suportados e o imposto sobre o rendimento;

- ROVendas (rentabilidade operacional das vendas): reflete quanto é que as vendas contribuem para o resultado operacional da empresa;
- RCP (rentabilidade do capital próprio): reflete a capacidade de gestão dos investimentos dos detentores de capital em gerar retorno financeiro;
- MRADGFI (margem de resultado antes de depreciações, gastos de financiamento e impostos): trata-se de um indicador que apresenta o que a empresa espera ganhar após a venda de um produto ou a prestação de um serviço sem ter em conta depreciações, gastos de financiamento e impostos;
- ML (margem de lucro): trata-se de um indicador que retrata o que a empresa espera ganhar após a venda de um produto ou a prestação de um serviço, o seu ganho líquido.

Quanto às variáveis de crescimento, são englobadas quatro, sendo que estas retratam a evolução entre dois períodos consecutivos, nomeadamente, da rentabilidade do capital próprio, do número de empregados, do valor do ativo e do volume de vendas.

Quanto às variáveis de endividamento, são englobadas sete:

- End (endividamento): retrata o nível de capital alheio utilizado para financiar as atividades da empresa. Este indicador procura avaliar o risco de incumprimento por parte da empresa;
- Solv (solvabilidade): retrata a capacidade dos capitais próprios de suportarem o nível de endividamento da empresa;
- AF (autonomia financeira): traduz a parte do ativo que está a ser financiada a partir dos capitais próprios da empresa;
- PCAtivo: traduz a percentagem de passivo corrente no total do ativo;
- FCDLP: traduz a capacidade da empresa em gerar fluxos suficientes da sua atividade operacional que lhe permita remunerar os capitais alheios de longo prazo;
- FCDLP: traduz a parte do fluxo de caixa que é gerado a partir de investimentos com dívida de longo prazo;
- DLPAtivo: traduz o nível de dívida de longo prazo utilizado para financiar as atividades da empresa;
- DLPCP: traduz o nível de dívida de longo prazo utilizado para o capital próprio da empresa.

Quanto às variáveis da estrutura do ativo são englobadas duas:

- ACAtivo: traduz a percentagem de ativo que é de curto prazo, isto é, que tem maior liquidez (inferior a 1 ano);
- FCAtivo: traduz o peso dos meios financeiros no total do ativo.

Quanto às variáveis de liquidez são englobadas quatro:

- LG (liquidez geral): traduz em que medida o endividamento de curto prazo se encontra coberto por ativos que poderão ser convertidos em meios financeiros também no curto prazo;
- ACPassivo: traduz em que medida o endividamento de curto e de longo prazo se encontra coberto por ativos que poderão ser convertidos em meios financeiros também no curto prazo;
- FMAtivo: representa a composição da empresa em termos de dívida de curto prazo comparativamente a meios de curto prazo. O objetivo de uma empresa é obter um fundo de maneio superior a zero por forma a ter mais ativos do que endividamento. A variável FMAtivo traduz a percentagem de ativo que é coberta pelo fundo de maneio;
- FCRO: traduz a parte dos fluxos de caixa que geram resultado operacional.

Quanto às variáveis de rotação são englobadas duas:

- VendasAtivo: traduz o número de vezes que o ativo é transformado em vendas num determinado intervalo de tempo;
- VendasAC: traduz o número de vezes que o ativo corrente (inferior a 1 ano) é transformado em vendas num determinado intervalo de tempo.

Adicionalmente, são analisadas as correlações lineares de Pearson entre as variáveis quantitativas tendo-se identificado quatro correlações muito fortes (Pearson > 0,9)[8]. No entanto, opta-se por não excluir nenhuma das variáveis apesar de alguma redundância que possa existir nos dados.

3.3. Modelação e avaliação

A Tabela 3.5 resume as técnicas utilizadas ao longo do projeto, quer para as etapas da compreensão e preparação dos dados, como para cada um dos objetivos específicos da investigação, isto é, para a etapa da modelação.

[8] Ver Anexo C – Correlações entre as variáveis independentes (página 65)

Tabela 3.5: Principais técnicas de analise de dados utilizadas no projeto analítico

Técnicas	Objetivos
Estatística descritiva univariada	Compreensão dos dados
Análise e tratamento de *outliers*	Preparação de dados
Estatística descritiva bivariada	(1)
Testes de hipóteses - teste-t e independência do qui-quadrado	(1)
Árvores de decisão para classificação	(2); (3)

Fonte: Elaboração própria

Para a concretização do primeiro objetivo específico, recorre-se a uma análise estatística bivariada, descritiva e inferencial, para comparar a situação financeira e não financeira entre empresas ativas e não ativas. Em particular, recorre-se ao teste-t para identificar diferenças significativas das médias das variáveis independentes quantitativas entre os dois *status*. Quando as variáveis independentes são qualitativas, recorre-se ao teste de independência do qui-quadrado. É importante destacar que o nível de significância (α) adotado nos testes estatísticos é de 0,05 (Jadhav e Shandilya, 2023). Adicionalmente às medidas descritivas tradicionais (e.g., média, desvio-padrão, percentagens) apresentam-se as medidas de associação Eta e V de Cramer (VC), que medem a intensidade da relação entre as duas variáveis em estudo. A intensidade da relação varia entre 0 e 1 sendo que, 0 a 0,2 traduz uma relação muito fraca; 0,2 a 0,4 uma relação fraca; 0, 4 a 0,7 uma relação moderada; 0,7 a 0,9 uma relação forte; 0,9 a 1 uma relação muito forte (Laureano, 2020).

Para a realização do segundo objetivo específico, são elaborados diversos modelos preditivos tendo em conta, de acordo com a RL, a técnica de árvores de decisão devido à sua elevada capacidade de perceção por parte de quem a interpreta (Olson *et al.*, 2012). Os modelos são considerados bons quando apresentam as métricas usuais superiores aos valores médios obtidos na RL (Tabela 3.1) visto que os autores, em todos os artigos analisados, consideram que obtiveram bons modelos.

Existem dois tipos de árvores de decisão: árvores de regressão e de classificação. Uma vez que a variável dependente é binária utilizam-se árvores de classificação (Wei-Yin Loh, 2008), mais especificamente três dos algoritmos mais adotados, nomeadamente *Classification and Regression Trees* (*CART*), *Chi-squared Automatic Interaction Detection* (*CHAID*) e *C5.0* (Chiang *et al.*, 2013), sendo que na RL apenas são identificados o *CART* e o *C5.0*. O algoritmo *CART* foi desenvolvido por Breiman *et al.* (1984), sendo um algoritmo de previsão que divide os dados tendo em conta a homogeneidade dos mesmos e o tipo de variável em questão (Tian

e Zhang, 2022). Para tal, sendo uma variável dependente qualitativa, é utilizado o índice de Gini para medir a dispersão dos dados (James *et al.*, 2021). O algoritmo *CHAID* foi desenvolvido por Kass (1980), que utiliza o teste de independência do Qui-Quadrado para identificar a divisão mais significativa avaliando a associação entre cada variável explicativa e a variável dependente (Park *et al.*, 2018). Por fim, o algoritmo *C5.0* foi desenvolvido por Quinlan *et al.* (1994) e utiliza como medida de avaliação o ganho de informação para selecionar o atributo mais informativo, para que seja efetuada a melhor divisão possível dos nós da árvore de decisão tendo em conda a característica mais informativa (Tian e Zhang, 2022).

Os modelos obtidos através destes algoritmos podem ser construídos recorrendo à combinação de modelos (usualmente designados por *ensembles*). Recorrendo ora à técnica de *bagging*, que visa melhorar a capacidade de generalização dos mesmos, aumentando a sua estabilidade, ou através da técnica *boosting*, que pretende melhorar a precisão do modelo na fase de treino (IBM, 2023). Para a criação desses mesmos modelos, é necessário determinar a profundidade máxima da árvore, isto é, o número máximo de níveis da mesma. Por defeito, está estipulado um valor de cinco níveis no caso do *CART* e do *CHAID*, sendo que no C5.0 são definidos automaticamente. Para simplicidade da análise, é mantida a profundidade estipulada no *software* utilizado. Adicionalmente, também é necessário determinar o número mínimo de casos por Nó Pai (por defeito, dois) e o número mínimo de casos por Nó Filho (por defeito, um).

Relativamente às parametrizações, e uma vez que as empresas ativas excedem as não ativas (Tabela 3.4 e **Error! Reference source not found.**), para evitar que o modelo classifique as empresas como ativas utiliza-se o balanceamento dos dados. Para equilibrar podem ser utilizadas duas técnicas: *boost* e *reduce*. Ao utilizar a técnica *boost* são replicados, aleatoriamente, alguns casos da classe minoritária; contrariamente a técnica *reduce*, retira da amostra, aleatoriamente, casos da classe maioritária (IBM, 2023). Assim, por forma a diminuir a dimensão dos dados e o tempo de processamento dos mesmos recorre-se ao *reduce*. A Tabela 3.6 resume as parametrizações de alguns dos modelos realizados[9] com os diferentes algoritmos, destacando-se que são utilizados diferentes anos de histórico para medir o impacto de mais informação na qualidade dos modelos.

[9] Foram realizados mais de 100 modelos com diferentes parametrizações, mas devido à extensão desta análise, a mesma não foi incluída nesta dissertação.

Tabela 3.6: Parametrização dos modelos preditivos

	Modelos										
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
Algoritmo	CART	CHAID	C5.0	CART	CHAID	C5.0	CHAID	CHAID	CART	CHAID	C5.0
Ensembles	-	-	-	*Boosting*	*Boosting*	*Boosting*	*Bagging*	*Boosting*	*Boosting*	*Boosting*	*Boosting*
Prof. máxima	5	5	5	5	5	5	5	5	5	5	5
Nº de casos Nó Pai	2	2	2	2	2	2	2	2	2	2	2
Nº de casos Nó Filho	1	1	1	1	1	1	1	1	1	1	1
Anos em análise	n-1	n-1	n-1	n-1	n-1	n-1	n-1	n-1	n-1 a n-8	n-1 a n-8	n-1 a n-8
Variáveis	Todas	Todas	Todas	Todas	Todas	Todas	Todas	Todas sem AF e ROA	Todas	Todas	Todas

Fonte: Elaboração própria

Notas: Nº: Número; Prof.: Profundidade

Para validar os modelos adota-se uma abordagem em que 70% das empresas são utilizadas como amostra de treino, enquanto os 30% restantes constituem a amostra de teste (Abrantes, 2020; Delen *et al.*, 2013). Para avaliar a qualidade dos modelos são escolhidas diversas métricas tendo em conta a matriz de classificação (Delen *et al.*, 2013) apresentada na Tabela 3.7, em que se considera a classe positiva as empresas não ativas (insolvente/falida).

Tabela 3.7: Matriz de classificação

Classe observada	**Classe prevista**	
	Ativa	**Não ativa**
Ativa	VN (Verdadeiro negativo)	FP (Falso positivo)
Não ativa	FN (Falso negativo)	VP (Verdadeiro positivo)

Fonte: Elaboração própria

Mais especificamente, recorre-se às métricas: *accuracy* (ou percentagem de casos corretamente classificados (PCCC)), especificidade, sensibilidade, precisão e *F1-Score*. A especificidade traduz a percentagem de empresas ativas corretamente classificadas, contrariamente a sensibilidade traduz a percentagem de empresas não ativas corretamente classificadas. A precisão traduz a percentagem de empresas classificadas como não ativas que são, efetivamente, não ativas. Já o *F1-Score* traduz o *trade-off* entre a sensibilidade e a precisão quando ambas as métricas são importantes para o problema em causa. A Tabela 3.8 resume as fórmulas utilizadas para o cálculo destas métricas.

Tabela 3.8: Métricas de avaliação de qualidade dos modelos de classificação

Métrica	Fórmula
Accuracy	$\frac{VP + VN}{VP + VN + FP + FN}$
Especificidade	$\frac{VN}{VN + FP}$
Sensibilidade	$\frac{VP}{VP + FN}$
Precisão	$\frac{VP}{VP + FP}$
F1-Score	$\frac{2 * Sensibilidade * Precisão}{Sensibilidade + Precisão}$

Fonte: Elaboração própria

Notas: FN: Falsos negativos; FP: Falsos positivos; VN: Verdadeiros negativos; VP: Verdadeiros positivos

Adicionalmente foi utilizada uma outra métrica, *Area Under the ROC Curve* (*AUC*), pois é considerada uma boa métrica para a avaliação de modelos com variáveis dependentes binárias. É importante referir que a *ROC (Receiver Operating Characteristic) curve* apresenta a relação entre a sensibilidade e a especificidade. O valor da *AUC* varia entre 0 e 1, sendo que o valor 0 significa que todas as previsões que o modelo efetuou estão erradas e o valor 1 corresponde a um modelo com previsões 100% corretas (Vanderlooy e Hüllermeier, 2008).

Por fim, o terceiro objetivo específico é identificar os perfis de empresas propensas à insolvência/falência ativas, através de uma análise dos nós finais da árvore de decisão que apresentam os resultados mais confiáveis (confiança acima de 80%) e suporte estatisticamente relevante (acima de 1.000 empresas).

4. Resultados e discussão

Este capítulo apresenta os resultados obtidos a partir da aplicação da metodologia adotada para cada um dos objetivos específicos definidos.

4.1. Caracterização da situação não financeira e financeira por *status* de empresa

A análise destes resultados (Tabela 4.1) permite verificar, no que respeita à situação não financeira, que o tipo de empresa coletiva predomina face ao unipessoal, tanto nas empresas ativas (72,40%) como nas empresas não ativas (64,02%), pelo que esta ligeira diferença se traduz numa relação significativa ($p < 0,001$) mas muito fraca (VC= 0,05).

No caso do setor, o predominante é o 7 - Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos, tanto nas empresas ativas (24,65%) como nas empresas não ativas (26,98%). Estas ligeiras diferenças traduzem-se numa relação significativa entre o setor e o *status* da empresa ($p < 0,001$), embora muito fraca (VC = 0,05).

Relativamente à região, o Norte predomina tanto nas empresas ativas (35,04%) como nas empresas não ativas (35,62%), sendo que a maior diferença entre os dois *status* das empresas se encontra na AML (Ativas- 29,44%; Não ativas- 33,67%). Desta forma, a variável região encontra-se significativamente relacionada com o *status* da empresa ($p < 0,001$) apesar desta ser muito fraca (VC = 0,04).

No caso da variável Dimensão, a dimensão das empresas predominante é a pequena empresa tanto nas empresas ativas (78,97%) como nas empresas não ativas (93,86%). Esta variável encontra-se significativamente relacionada com o *status* da empresa ($p < 0,001$) apesar de com uma relação muito fraca (VC = 0,11).

A forma jurídica predominante é a sociedade por quotas de responsabilidade limitada tanto nas empresas ativas (93,79%) como nas empresas não ativas (96,70%). Esta variável encontra-se significativamente relacionada com o *status* da empresa ($p < 0,001$), apesar de com uma relação muito fraca (VC = 0,04).

Tabela 4.1: Distribuição das variáveis demográficas (qualitativas) por *status* das empresas

Variáveis	Descrição	*Status*				Teste Qui-Quadrado (χ^2) V de Cramer (VC)
		Ativa		Não ativa		
		Nº	%	Nº	%	
Unipessoal	Coletiva	465.511	72,40	41.168	64,02	χ^2 = 2.021,34; p < 0,001 VC = 0,05; p < 0,001
	Unipessoal	177.472	27,60	23.140	35,98	
Setor	1	26.173	4,07	1.266	1,97	χ^2 = 2.919,31; p < 0,001 VC = 0,06; p < 0,001
	2	1.247	0,19	79	0,12	
	3	69.094	10,75	6.308	9,81	
	4	1.370	0,21	44	0,07	
	5	1.615	0,25	132	0,21	
	6	67.007	10,42	7.515	11,69	
	7	158.505	24,65	17.353	26,98	
	8	32.777	5,10	3.392	5,27	
	9	51.931	8,08	5.192	8,07	
	10	16.574	2,58	2.550	3,97	
	11	11.612	1,81	629	0,98	
	12	47.569	7,40	3.051	4,74	
	13	67.395	10,48	7.064	10,98	
	14	20.670	3,21	2.728	4,24	
	15	16	0,00	4	0,01	
	16	7.754	1,21	1.061	1,65	
	17	43.361	6,74	3.447	5,36	
	18	8.792	1,37	1.145	1,78	
	19	9.521	1,48	1.348	2,10	
Região	Norte	225.280	35,04	22.907	35,62	χ^2 = 981,63; p < 0,001 VC = 0,04; p < 0,001
	Centro	135.353	21,05	11.517	17,91	
	AML	189.301	29,44	21.653	33,67	
	Alentejo	40.118	6,24	3.143	4,89	
	Algarve	29.841	4,64	2.641	4,11	
	Madeira	14.149	2,20	1.779	2,77	
	Açores	8.941	1,39	668	1,04	
Dimensão	P	507.739	78,97	60.358	93,86	χ^2 = 8.242,56; p < 0,001 VC = 0,11; p < 0,001
	M	117.857	18,33	3.672	5,71	
	G	15.318	2,38	262	0,41	
	MG	2.069	0,32	16	0,02	
FormaJuridica	SA	39.938	6,21	2.123	3,30	χ^2 = 885,15; p < 0,001 VC = 0,04; p < 0,001
	LDA	603.045	93,79	62.185	96,70	

Fonte: Elaboração própria

Notas: AML: Área Metropolitana de Lisboa; G: Grande; LDA: Sociedade por quotas de responsabilidade limitada; M: Média; MG; Muito grande; Nº: número de empresas; P: Pequena; SA: Sociedade Anónima; Informação acerca dos setores presente na Tabela 3.2.

Já no que respeita à antiguidade a Tabela 4.2 permite verificar que existe uma diferença significativa entre a média dos anos de atividade das empresas ativas (15,23 anos) e a das não ativas (12,69 anos) ($p < 0,001$), apesar de existir uma relação muito fraca entre as duas características (Eta = 0,09).

Por fim, quanto à composição do órgão de gestão verifica-se que existe uma diferença significativa entre a média da percentagem do género feminino no total do conselho de administração das empresas ativas (24,00%) e das não ativas (19,00%) ($p < 0,001$), apesar de ter uma relação muito fraca (Eta = 0,04).

Tabela 4.2: Distribuição das variáveis demográficas (quantitativas) por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
Anos_atividade	Ativa	642.983	15,23	13,37	2	12	300	t = 49,08; p < 0,001 Eta = 0,09
	Não ativa	64.308	12,69	12,45	2	8	140	
Proporcao_Feminina	Ativa	36.894	0,24	0,31	0	0	1	t = 3,48; p < 0,001 Eta = 0,04
	Não ativa	406	0,19	0,31	0	0	1	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas

A partir da Tabela 4.3, verifica-se que, em termos de rentabilidade, existem diferenças significativas entre os *status* das empresas, apesar da relação ser fraca (Eta = 0,40), à exceção da rentabilidade do capital próprio em que não existe uma diferença significativa entre a média do rácio das empresas ativas (0,13) e das não ativas (0,12) ($p = 0,09 > \alpha = 0,05$), sendo que apresenta uma relação moderada (Eta = 0,43). Ou seja, em média as empresas não ativas apresentam uma rentabilidade negativa e as empresas ativas apresentam uma rentabilidade positiva.

Quanto às margens, existem diferenças significativas entre os dois *status* das empresas ainda que com uma relação fraca (Eta = 0,27). Assim, as empresas ativas apresentam, tendencialmente, margens positivas e as empresas não ativas apresentam, em média, rentabilidades negativas ou muito próximas de zero.

Tabela 4.3: Distribuição das variáveis de rentabilidade por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
ROAtivo_n1	Ativa	621.850	0,02	0,33	-2,33	0,04	0,79	t = 81,00; p < 0,001 Eta = 0,40
	Não ativa	63.639	-0,18	0,61	-2,33	-0,01	0,79	
RLAtivo_n1	Ativa	621.876	0,00	0,32	-2,36	0,02	0,67	t = 81,90; p < 0,001 Eta = 0,39
	Não ativa	63.639	-0,21	0,61	-2,36	-0,01	0,67	
ROVendas_n1	Ativa	589.339	-0,01	0,58	-4,58	0,04	1,04	t = 67,29; p < 0,001 Eta = 0,36
	Não ativa	52.661	-0,32	1,03	-4,58	0,00	1,04	
RCP_n1	Ativa	618.263	0,13	1,08	-6,14	0,08	5,26	t = 1,68; **p = 0,090** Eta = 0,43
	Não ativa	60.322	0,12	1,60	-6,14	0,07	5,26	
MRADGFI_n1	Ativa	587.945	0,12	0,25	-0,77	0,08	0,98	t = 63,63; p < 0,001 Eta = 0,27
	Não ativa	48.898	0,02	0,33	-0,77	0,03	0,98	
ML_n1	Ativa	583.771	0,05	0,24	-0,82	0,03	0,94	t = 51,65; p < 0,001 Eta = 0,27
	Não ativa	48.117	-0,02	0,32	-0,82	0,01	0,94	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas; Negrito: diferenças não significativas

No caso das variáveis de crescimento (Tabela 4.4), a análise destes resultados permite verificar que existem diferenças significativas entre os dois *status* das empresas sendo mais notório na variável de crescimento da rentabilidade do capital próprio, apresentando uma relação moderada (Eta = 0,54). Assim, em média, as empresas ativas tendem a ter variações positivas e mais elevadas do que as empresas não ativas, à exceção da rentabilidade do capital próprio que se tratou de uma variação negativa em ambos os *status* das empresas, mas ainda assim com diferença significativa entre ambos.

Tabela 4.4: Distribuição das variáveis de crescimento por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
Var.RCP_n1	Ativa	492.496	-0,43	11,85	-75,23	-0,38	62,39	t = 9,12; p < 0,001 Eta = 0,54
	Não ativa	51.942	-1,11	16,49	-75,23	-0,69	62,39	
Var.Empr._n1	Ativa	555.123	0,07	0,40	-1,00	0,00	2,00	t = 65,53; p < 0,001 Eta = 0,20
	Não ativa	45.908	-0,09	0,48	-1,00	0,00	2,00	
Var.ativo_n1	Ativa	500.541	0,28	1,15	-0,77	0,03	9,09	t = 31,71; p < 0,001 Eta = 0,40
	Não ativa	58.229	0,12	1,20	-0,77	-0,05	9,09	
Var.Vendas_n1	Ativa	472.122	0,33	1,50	-1,00	0,03	11,68	t = 22,68; p < 0,001 Eta = 0,42
	Não ativa	51.096	0,14	1,79	-1,00	-0,15	11,68	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas

No caso das variáveis de endividamento (Tabela 4.5), a análise destes resultados permite verificar que existem diferenças significativas entre os dois *status* das empresas sendo mais notório na variável FCDLP_n1 em que apresenta uma relação moderada (Eta = 0,55). Neste rácio, tendencialmente, as empresas ativas apresentam valores superiores a 1 e as empresas não ativas apresentam valores inferiores a 1.

Tabela 4.5: Distribuição das variáveis de endividamento por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
End_n1	Ativa	629.403	0,67	0,6	0,00	0,62	4,78	t = -56,44; p < 0,001 Eta = 0,33
	Não ativa	63.639	0,90	1,04	0,00	0,70	4,78	
Solv_n1	Ativa	620.385	4,14	14,9	-0,79	0,60	125,30	t = -27,16; p < 0,001 Eta = 0,47
	Não ativa	60.391	6,60	21,75	-0,79	0,35	125,30	
AF_n1	Ativa	620.385	0,32	0,59	-3,75	0,37	0,99	t = 64,51; p < 0,001 Eta = 0,32
	Não ativa	60.391	0,05	1,04	-3,75	0,26	0,99	
PCAtivo_n1	Ativa	626.121	0,42	0,44	0,00	0,31	3,15	t = -54,78; p < 0,001 Eta = 0,29
	Não ativa	63.639	0,58	0,71	0,00	0,36	3,15	
FCDLP_n1	Ativa	296.842	1,51	6,42	-7,64	0,21	51,46	t = 26,95; p < 0,001 Eta = 0,55
	Não ativa	20.391	0,37	5,82	-7,64	-0,03	51,46	
DLPAtivo_n1	Ativa	409.114	0,24	0,30	0,00	0,13	1,67	t = 57,98; p < 0,001 Eta = 0,34
	Não ativa	63.639	0,15	0,34	0,00	0,00	1,67	
DLPCP_n1	Ativa	408.162	1,07	5,09	-17,33	0,17	33,42	t = 31,73; p < 0,001 Eta = 0,48
	Não ativa	60.322	0,38	3,98	-17,33	0,00	33,42	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas

No caso das variáveis de estrutura do ativo (Tabela 4.6), a análise destes resultados permite verificar que existem diferenças significativas entre os dois *status* das empresas sendo mais notório na variável FCAtivo_n1 em que apresenta uma relação moderada (Eta = 0,40). Neste rácio, tendencialmente, as empresas ativas apresentam valores positivos enquanto as empresas não ativas apresentam valores negativos.

Tabela 4.6: Distribuição das variáveis da estrutura do ativo por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
ACAtivo_n1	Ativa	626.121	0,71	0,29	0,02	0,81	1,00	t = -84,14; p < 0,001 Eta = 0,22
	Não ativa	63.639	0,81	0,28	0,02	0,97	1,00	
FCAtivo_n1	Ativa	621.859	0,04	0,32	-2,27	0,05	0,73	t = 85,41; p < 0,001 Eta = 0,40
	Não ativa	63.639	-0,17	0,60	-2,27	0,00	0,73	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas

No caso das variáveis de liquidez (Tabela 4.7), a análise destes resultados permite verificar que existem diferenças significativas entre os dois *status* das empresas sendo mais notório na variável LG_n1 em que apresenta uma relação moderada (Eta = 0,51). Neste rácio, tendencialmente, as empresas ativas apresentam valores, em média, inferiores a 10, enquanto as empresas não ativas apresentam valores, em média, superiores a 10.

Tabela 4.7: Distribuição das variáveis de liquidez por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
LG_n1	Ativa	616.245	9,68	32,46	0,05	2,04	271,34	t = -23,93; p < 0,001 Eta = 0,51
	Não ativa	59.114	14,10	43,60	0,05	1,80	271,34	
ACPassivo_n1	Ativa	620.385	3,63	10,58	0,02	1,17	89,93	t = -33,10; p < 0,001 Eta = 0,43
	Não ativa	60.419	5,84	16,10	0,02	1,11	89,93	
FMAtivo_n1	Ativa	577.510	0,22	0,31	-0,67	0,17	0,97	t = 48,28; p < 0,001 Eta = 0,45
	Não ativa	63.639	0,15	0,34	-0,67	0,03	0,97	
FCRO_n1	Ativa	586.993	0,09	0,23	-0,78	0,06	0,94	t = 64,91; p < 0,001 Eta = 0,27
	Não ativa	48.671	-0,01	0,31	-0,78	0,02	0,94	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas

No caso das variáveis de rotação (Tabela 4.8), a análise destes resultados permite verificar que apenas existe uma diferença significativa entre os dois *status* das empresas na variável VendasAtivo_n1, apesar de com uma relação fraca (Eta = 0,35). Neste rácio, tendencialmente, as empresas ativas apresentam valores, em média, inferiores a 1,4, enquanto as empresas não ativas apresentam valores, em média, superiores a 1,4.

Tabela 4.8: Distribuição das variáveis de rotação por *status* das empresas

Variáveis	*Status*	RV	Média	DP	Mínimo	Mediana	Máximo	Teste-t Eta
VendasAtivo_n1	Ativa	622.822	1,39	1,57	0,00	0,97	10,06	t = -12,47; p < 0,001 Eta = 0,35
	Não ativa	63.639	1,49	2,12	0,00	0,77	10,06	
VendasAC_n1	Ativa	622.406	2,38	3,16	0,00	1,47	21,35	t = 0,77; **p = 0,440** Eta = 0,40
	Não ativa	63.373	2,36	3,96	0,00	1,03	21,35	

Fonte: Elaboração própria

Notas: DP: Desvio-padrão; RV: respostas válidas; Negrito: diferenças não significativas

Com os resultados obtidos e analisados é possível caracterizar as situações não financeira (demográfica) e financeira de cada um dos tipos de empresas.

As empresas ativas tendem a ser sociedades por quotas de responsabilidade limitada (93,79%) e coletivas (72,40%), de pequena dimensão (78,97%), localizadas no Norte (35,04%) e pertencentes ao setor do comércio por grosso e a retalho; reparação de veículos automóveis e motociclos (24,65%). São empresas com alguma antiguidade (M = 15,23 anos) e que, em média, em cada quatro elementos do conselho de administração um é do sexo feminino (M = 24,00%).

Analisando agora a sua situação financeira, esta caracteriza-se por, em média e em geral, uma rentabilidade positiva (por exemplo, ROAtivo_n1 = 0,02), e um crescimento anual, nomeadamente ao nível da rentabilidade do capital próprio, em média, de -43,00%. Caracterizam-se também por apresentarem, em média, um rácio de fluxo de caixa pela dívida de longo-prazo superior a 1 e um rácio do fluxo de caixa pelo ativo, em geral, positivo. Adicionalmente, as empresas ativas caracterizam-se por uma liquidez geral, em média, inferior a 10 e um rácio de vendas pelo ativo, tendencialmente, inferiores a 1,4.

Já as empresas não ativas tendem a ser sociedades por quotas de responsabilidade limitada (96,70%) e coletivas (64,02%), de pequena dimensão (93,86%), localizadas no Norte (35,62%) e pertencentes ao setor do comércio por grosso e a retalho; reparação de veículos automóveis e motociclos (26,98%). São empresas com alguma antiguidade (M = 12,69 anos) e que, em média, cada cinco elementos do conselho de administração um é do sexo feminino (M = 19,00%).

Analisando agora a sua situação financeira, esta caracteriza-se por, em média e em geral, uma rentabilidade negativa (por exemplo, ROAtivo_n1 = -0,18), e um crescimento anual, nomeadamente ao nível da rentabilidade do capital próprio, em média, de -111,00%. Caracterizam-se também por apresentarem, em média, um rácio de fluxo de caixa pela dívida de longo-prazo inferior a 1 e um rácio do fluxo de caixa pelo ativo, em geral, negativo. Adicionalmente, as empresas não ativas caracterizam-se por uma liquidez geral, em média, superior a 10 e um rácio de vendas pelo ativo, tendencialmente, superior a 1,4.

4.2. Modelo preditivo das insolvências/falências

Os diferentes modelos estimados, tendo em conta os parâmetros da Tabela 3.6, são apresentados na Tabela 4.9, tanto para a amostra de treino, quer para a amostra de teste. Os resultados evidenciam que os algoritmos que utilizam o *boosting* apresentam melhores resultados para o *CART* e *CHAID*, mais especificamente, o modelo com melhores resultados foi o modelo J que incluiu todas as variáveis e ainda o intervalo de tempo de n-1 a n-8 (*Accuracy*: 90,53%; Precisão: 49,05%; Sensibilidade: 91,06%; Especificidade: 90,47%; *F1-Score*: 0,64; *AUC*: 0,97). Porém, este modelo é um modelo muito complexo e, como tal, opta-se por selecionar um modelo mais simples e que apresente, igualmente, boas métricas na amostra de teste: o modelo E (*Accuracy*: 82,20%; Precisão: 31,85%; Sensibilidade: 82,97%; Especificidade: 82,12%; *F1-Score*: 0,46; *AUC*: 0,91).

Tabela 4.9: Resultados dos modelos preditivos das insolvências/falências

		Modelos										
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
	Algoritmo	*CART*	*CHAID*	*C5.0*	*CART*	*CHAID*	*C5.0*	*CHAID*	*CHAID*	*CART*	*CHAID*	*C5.0*
	Ensembles	-	-	-	*Boosting*	*Boosting*	*Boosting*	*Bagging*	*Boosting*	*Boosting*	*Boosting*	*Boosting*
Treino	***Accuracy* (PCCC)**	71,10%	74,53%	78,23%	76,92%	83,05%	86,12%	80,46%	83,17%	79,53%	91,13%	82,48%
	Precisão	71,23%	84,16%	78,79%	78,53%	82,51%	85,33%	78,55%	83,06%	77,88%	90,72%	85,05%
	Sensibilidade	70,62%	76,68%	77,32%	73,95%	83,89%	87,16%	83,72%	83,29%	82,33%	91,62%	78,73%
	Especificidade	71,57%	80,25%	79,14%	79,87%	82,22%	85,08%	77,22%	83,05%	76,75%	90,65%	86,21%
	F1-Score	0,71	0,79	0,78	0,76	0,83	0,86	0,81	0,83	0,80	0,91	0,82
	AUC	0,74	0,88	0,83	0,82	0,91	0,93	0,89	0,91	0,86	0,97	0,90
Teste	***Accuracy* (PCCC)**	71,58%	74,58%	78,17%	79,21%	82,20%	84,23%	77,73%	82,64%	77,01%	90,53%	85,30%
	Precisão	20,01%	83,87%	25,66%	26,76%	31,85%	34,40%	26,91%	32,37%	26,06%	49,05%	35,83%
	Sensibilidade	70,25%	24,94%	73,06%	73,26%	82,97%	79,74%	83,56%	82,44%	82,31%	91,06%	76,70%
	Especificidade	71,71%	38,45%	78,68%	79,81%	82,12%	84,69%	77,15%	82,66%	76,47%	90,47%	86,17%
	F1-Score	0,31	0,75	0,38	0,39	0,46	0,48	0,41	0,46	0,40	0,64	0,49
	AUC	0,74	0,88	0,81	0,82	0,91	0,90	0,89	0,91	0,86	0,97	0,89

Fonte: Elaboração própria

As métricas mais importantes para avaliar os modelos, tendo em conta o foco deste estudo em classificar empresas não ativas, são a *accuracy* e a sensibilidade. Isto deve-se ao facto de a *accuracy* indicar a percentagem total de casos corretamente identificados e a sensibilidade traduzir a percentagem de casos corretamente classificados das empresas não ativas, visto este último ser o foco desta investigação.

Adicionalmente, dentro das variáveis que apresentam maior importância na predição dos modelos, as que surgem com maior frequência são o rácio da autonomia financeira (AF_n1 – sete modelos) e o rácio do fluxo de caixa/ativo (FCAtivo_n1 – sete modelos). Através da análise da importância das 13 variáveis mais importantes (de um total de 202 variáveis consideradas como potencialmente explicativas do *status*) associadas ao modelo E (Figura 4.1), constata-se que das seis variáveis de rentabilidade, cinco surgem como variáveis importantes para a previsão de insolvências/falências. Surge ainda uma variável demográfica, a dimensão da empresa, as variáveis da estrutura do ativo, três das sete variáveis de endividamento e ainda uma variável de liquidez (FCRO_n1).

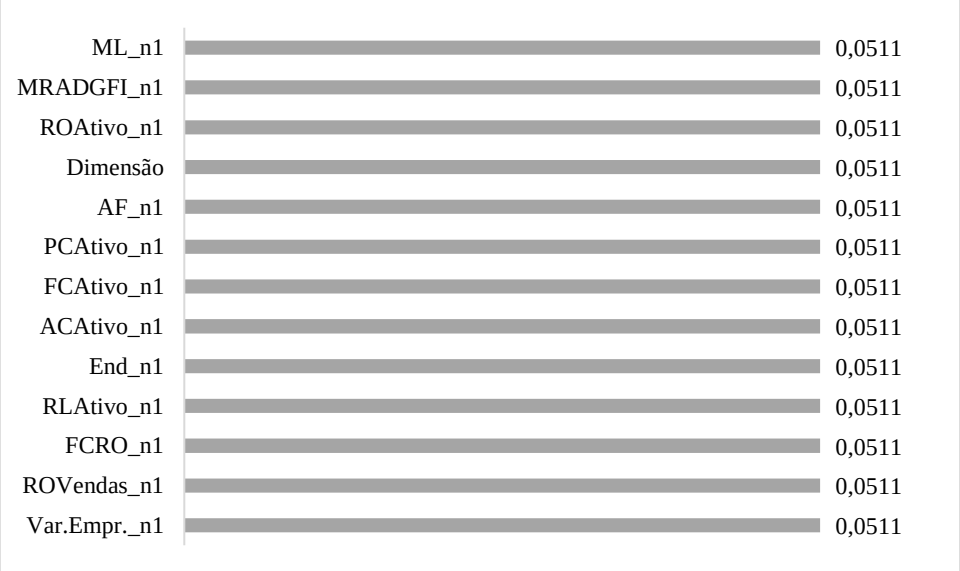

Figura 4.1: Importância das principais variáveis preditoras do modelo E

Fonte: Elaboração própria

4.3. Perfis associados às empresas ativas e não ativas

De forma a identificar perfis de empresas ativas e não ativas, recorre-se à análise dos nós terminais com maior confiança e suporte da árvore de decisão proveniente do modelo B. Os melhores modelos preditivos são encontrados utilizando o *boosting*, porém não é possível

aceder às regras da árvore para se alcançar os perfis das empresas visto que este tipo de modelo resulta de uma combinação de modelos (10 árvores). Como tal, é utilizado o modelo mais simples e com melhores métricas para se descrever os perfis das empresas, seguindo o algoritmo selecionado como o melhor (CHAID).

Assim sendo, alguns perfis relacionados com as empresas ativas podem ser obtidos pela análise das seguintes regras de decisão:

- Se DLPAtivo_n1 > 0 e DLPAtivo_n1 <= 0,066, e FCAtivo_n1 > 0,001 e FCAtivo_n1 <= 0,089, e Anos_atividade > 13 então *status* = ativa (nó 58; suporte = 1.423 e confiança de 81,6%);
- Se DLPAtivo_n1 > 0,066 e DLPAtivo_n1 <= 0,180, e FCAtivo_n1 > 0,051, e Anos_atividade > 9 então *status* = ativa (nó 60; suporte = 2.093 e confiança de 85,8%).

Já para as empresas não ativas, o perfil relacionado com estas tende a ser:

- Se DLPAtivo_n1 <= 0, e Var.Empr._n1 > -0,500 e Var.Empr._n1 <= -0,001, e FCAtivo_n1 <= -0,027 então *status* = não ativa (nó 40; suporte = 1.134 e confiança de 85,90%);
- Se DLPAtivo_n1 <= 0, e Var.Empr._n1 > -0,001 e Var.Empr._n1 <= 0, e Var.Vendas_n1 <= -0,789 então *status* = não ativa (nó 42; suporte = 1.272 e confiança de 95,20%).

4.4. Discussão dos resultados

Dos resultados obtidos, pode concluir-se que efetivamente a autonomia financeira e o rácio do fluxo de caixa/ativo são variáveis importantes para diferenciar empresas ativas das não ativas, ou seja, para prever a passagem à inatividade das empresas (Antulov-Fantulin *et al.*, 2021).

Em termos das variáveis não financeiras, a inclusão da variável proporção feminina no conselho de administração também apresentou o resultado esperado, visto que de acordo com Antulov-Fantulin *et al.* (2021) esta não se revela importante para prever as insolvências/falências, sendo considerada a última variável em termos de importância para a previsão. Aliás, em termos de métricas de avaliação, observando um modelo preditivo sem a

inclusão desta variável, verificou-se que apenas teve uma diminuição das métricas, em média, de 0,3 pontos percentuais[10].

Por outro lado, a inclusão da técnica do *boosting* foi importante, uma vez que aumenta consideravelmente a qualidade dos modelos, isto é, as métricas de avaliação dos modelos preditivos, tal como sugerido pelo autor Lee *et al.* (2020). Desta forma, o modelo selecionado (E) consegue obter um melhor desempenho que os modelos identificados na RL e tal pode justificar-se devido à dimensão da amostra (grande) e ao contexto específico de Portugal.

Globalmente, e tendo em conta outras investigações que estudam o tema da predição das insolvências e falências tendo como preditores indicadores financeiros e não financeiros, pode concluir-se que o objetivo da criação de um modelo preditivo na presente investigação é alcançado visto que se conseguiram obter métricas superiores às obtidas, em média, na RL.

[10] Ver Anexo E – Comparação do modelo preditivo tendo em conta a variável da proporção feminina (página 70)

5. Conclusões

5.1. Sumário da investigação

Em Portugal, desde a crise financeira de 2008, o número de insolvências e falências de empresas aumentou moderadamente e, desde então, estes valores têm-se mantido. Estas insolvências e falências têm impactos irreversíveis, não só nas empresas que entram em inatividade, mas também nos diferentes *stakeholders*. De uma forma geral, a inatividade das empresas afeta a economia nacional na sua totalidade.

Tendo em conta a questão de investigação: "De que modo a estrutura financeira e societária de uma empresa permite prever a sua entrada em processos de insolvência e falências?", e o objetivo da mesma, o presente estudo identifica uma relação entre os indicadores financeiros e demográficos e o *status* das empresas. Além disso, a RSL realizada ajudou a concretizar com sucesso o mesmo. Com o intuito de identificar esta relação, e tendo em conta os três objetivos específicos, recorreu-se a técnicas de análise de dados, tais como, análise bivariada e árvores de decisão para identificar situações financeiras diferentes entre os dois tipos de empresas e prever o *status* das mesmas bem como determinar perfis de empresas associados às insolvências e falências mas também perfis de empresas ativas. Por fim, com base nos resultados obtidos, pode concluir-se que os objetivos delineados para este estudo foram cumpridos: os indicadores financeiros e não financeiros encontram-se relacionados com o *status* das empresas.

Dos rácios calculados, os mais importantes relacionam-se com a autonomia financeira e o fluxo de caixa/ativo que são as variáveis mais importantes na predição das insolvências/falências, relevantes em sete dos modelos apresentados.

Neste sentido, importa concluir que os indicadores financeiros e a estrutura societária das empresas contribuem de forma relevante para a predição do *status* das empresas. Adicionalmente, disponibilizam informação importante para se preverem as insolvências/falências das empresas, bastando para tal ter informação apenas do ano anterior, isto é, com base na situação financeira e não financeira de um ano é possível prever, recorrendo a uma árvore de decisão, com elevada taxa de acerto se uma empresa no ano seguinte vai entrar em insolvência/falência ou não. Desta forma, responde-se claramente à questão de investigação: "De que modo a estrutura financeira e societária de uma empresa permite prever a sua entrada em processos de insolvência e falências?".

5.2. Contributos

A resposta à questão de investigação e a concretização dos objetivos permite aos gestores conhecer as causas, financeiras e não financeiras, que podem levar uma empresa a entrar em dificuldades e, no limite, a um processo de insolvência ou de falência. De facto, uma boa previsão permite alertar para problemas de viabilidade, de modo que estes consigam delinear estratégias para evitar problemas financeiros e calcular as suas estimativas contabilísticas, aumentando a viabilidade e consistência das mesmas, evitando manipulação dos resultados obtidos.

Para os investidores, credores e bancos, entre outros, estes modelos permitem-lhes identificar as empresas com propensão para entrar em incumprimento, o que lhes permite tomar decisões de investimento e financiamento mais conscientes. Adicionalmente, conhecem os indicadores financeiros que mais comprometem a sustentabilidade financeira de uma empresa e os valores a partir dos quais devem alertar para uma mudança estratégica na empresa.

Aos auditores, os modelos permitem-lhes conhecer as empresas que têm mais propensão a incumprimento. O que os alerta para as empresas em que possa ser necessário realizar um trabalho mais profundo e detalhado de técnicas de auditoria de modo a evitarem qualquer problema.

Para o conhecimento científico e, em particular, para a área da contabilidade, foi publicado um artigo científico da minha autoria em *proceeedings* após apresentação do mesmo numa conferência sobre tecnologias aplicadas à contabilidade e auditoria, indexado no *WoS* e no *Scopus*, acerca de previsão de insolvências (Ildefonso et al., 2023). Além do mais, esta investigação, no seu capítulo dedicado à RSL, permite aumentar o grau de sistematização sobre temas de predição de dificuldades financeiras, insolvências e falências. Adicionalmente, a presente investigação evidencia que a junção das áreas científicas de contabilidade e de *business analytics* é cada vez mais relevante em termos académicos. Demonstração disso reside na possibilidade deste estudo ser utilizado como evidência empírica apresentada em aulas uma vez que os professores de métodos quantitativos aplicados à contabilidade têm neste estudo um exemplo, de Portugal, que evidencia a utilidade das técnicas analíticas para o estudo da contabilidade.

5.3. Limitações e pistas de investigação futuras

Importa referir que a falta de dados das empresas é a principal limitação desta investigação, dado que pode ter implicações no resultado da previsão, pois não se consegue ter a certeza se a

falta de valor em determinada rúbrica da empresa é fator para se dizer que uma empresa é ativa ou não ativa. No entanto esta limitação é de certa forma comum a todos os estudos que recorrem a dados secundários, facultados por bases de dados de subscrição, tal como a ORBIS Europe. Além disso, não sendo possível obter dados de empresas em dificuldades financeiras, torna a comparação com alguns estudos previamente realizados mais complicada.

No que diz respeito a investigações futuras e de forma a colmatar as limitações evidenciadas, é importante realçar que já se conseguiram modelos com elevada capacidade preditiva. Contudo ainda existe espaço para melhorar, através da inclusão de mais indicadores não financeiros que considerem a sustentabilidade ambiental (tema amplamente discutido na atualidade) ou alargar os estudos a outros países. Além do mais, é importante criar modelos preditivos que tenham a capacidade de perceber exatamente em que ano, caso aplicável, a empresa vai entrar em insolvência/falência, com a finalidade de obter resultados mais precisos. Adicionalmente, seria importante conseguir prever empresas portuguesas com dificuldades financeiras (Gutierrez *et al.*, 2020), para conseguir estratégias de recuperação menos complexas dado o estado de debilidade financeira menos avançado.

Além disso, seria interessante serem incluídas mais características do conselho de administração para além do género dos indivíduos, como por exemplo, as habilitações académicas e a experiência profissional.

Referências bibliográficas

Abrantes, C. (2020). Os modelos preditivos do sucesso de candidaturas a fundos europeus: o papel da manipulação de resultados. [Dissertação de mestrado, Iscte – Instituto Universitário de Lisboa]. Repositório Iscte. http://hdl.handle.net/10071/21827

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*. https://doi.org/10.1016/j.artmed.2022.102289

Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior and Organization*, *183*, 681–699. https://doi.org/10.1016/j.jebo.2021.01.014

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Empirical Research in Accounting: Selected Studies*, *4*, 71–111. https://doi.org/https://doi.org/10.2307/2490171

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., Mezzogori, D., & Zammori, F. (2022). Machine-learning models for bankruptcy prediction do industrial variables matter. *Spatial Economic Analysis*, *17*(2), 156–177. https://doi.org/https://doi.org/10.1080/17421772.2021.1977377

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees* (1.ª ed.). Taylor & Francis Group.

Cathcart, L., Dufour, A., Rossi, L., & Varotto, S. (2020). The differential impact of leverage on the default risk of small and large firms. *Journal of Corporate Finance*, *60*. https://doi.org/10.1016/j.jcorpfin.2019.101541

Chen, W. sen, & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, *36*(2 PART 2), 4075–4086. https://doi.org/10.1016/j.eswa.2008.03.020

Chiang, H. J., Tseng, C. C., & Torng, C. C. (2013). A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm. *Journal of Dental Sciences*, *8*(3), 248–255. https://doi.org/10.1016/j.jds.2013.04.009

Código da Insolvência e da Recuperação de Empresas atualizado - DL n.º 53/2004, de 18 de março. Disponível em https://dre.pt/dre/detalhe/decreto-lei/53-2004-538423 (acedido a 23/10/2022)

Código das Sociedades Comerciais atualizado - DL n.º 262/86, de 02 de setembro. Disponível em https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=524&tabela=leis (acedido a 01/08/2023)

Cooper, E., & Uzun, H. (2019). Corporate social responsibility and bankruptcy. *Studies in Economics and Finance*, *36*(2), 130–153. https://doi.org/10.1108/SEF-01-2018-0013

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*. https://doi.org/10.1016/j.dajour.2023.100164

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, *40*(10), 3970–3983. https://doi.org/10.1016/j.eswa.2013.01.012

Divsalar, M., Roodsaz, H., Vahdatinia, F., Norouzzadeh, G., & Behrooz, A. H. (2012). A robust data-mining approach to bankruptcy prediction. *Journal of Forecasting*, *31*(6), 504–523. https://doi.org/10.1002/for.1232

Esteban, S. A., Urquía-Grande, E., de Silva, A. M., & Pérez-Estébanez, R. (2022). Big Data, Accounting and International Development: Trends and challenges. *Cuadernos de Gestion*, *22*(1), 193–213. https://doi.org/10.5295/CDG.211513SA

Foster, D. P., & Stine, R. A. (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, *99*(466), 303–313. https://doi.org/10.1198/016214504000000287

Fundação Francisco Manuel dos Santos. https://www.ffms.pt/crises-na-economia-portuguesa/5046/filha-da-crise-financeira-internacional (Varejão, J., 2020) (acedido a 23/08/2022)

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), 236–247. https://doi.org/10.1016/j.ejor.2014.08.016

Gutierrez, E., Krupa, J., Minutti-Meza, M., & Vulcheva, M. (2020). Do going concern opinions provide incremental information to predict corporate defaults? *Review of Accounting Studies*, *25*(4), 1344–1381. https://doi.org/10.1007/s11142-020-09544-x

Huang, Y. P., & Yen, M. F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing Journal*, *83*. https://doi.org/10.1016/j.asoc.2019.105663

IBM - International Business Machines Corporation. https://www.ibm.com/docs/pt/spss-modeler/18.3.0?topic=nodes-balance-node (acedido a 24/07/2023)

Ildefonso, M. V. S., Laureano, R. M. S., & Vasarhelyi, M. A. (2023). Modelos preditivos de insolvências: uma revisão sistemática da literatura. CISTI'2023 – 18ª Conferência Ibérica de Sistemas e Tecnologias de Informação. https://doi.org/10.23919/CISTI58278.2023.10211516.

INE – Instituto Nacional de Estatística. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=599225312&DESTAQUESmodo=2 (acedido a 23/07/2023)

Isayas, Y. N. (2021). Financial distress and its determinants: Evidence from insurance companies in Ethiopia. *Cogent Business and Management*, *8*(1). https://doi.org/10.1080/23311975.2021.1951110

Jadhav, A., & Shandilya, S. K. (2023). Reliable Machine Learning Models for Estimating Effective Software Development Efforts: A Comparative Analysis. *Journal of Engineering Research*, 100150. https://doi.org/10.1016/j.jer.2023.100150

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(2), 119–127.

Kim, H., Cho, H., & Ryu, D. (2022). Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data. *Computational Economics*, *59*(3), 1231–1249. https://doi.org/10.1007/s10614-021-10126-5

Kwak, W., Shi, Y., Cheh, J. J., & Lee, H. (2004). Multiple Criteria Linear Programming Data Mining Approach: An Application for Bankruptcy Prediction. *Data Mining and Knowledge Management*, *3327*, 164–173. https://doi.org/https://doi.org/10.1007/978-3-540-30537-8_18

Lahmiri, S., & Bekiros, S. (2019). Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design. *Quantitative Finance*, *19*(9), 1569–1577. https://doi.org/https://doi.org/10.1080/14697688.2019.1588468

Laureano, R. (2020). Testes de Hipóteses e Regressão: O Meu Manual de Consulta Rápida (2ª Ed.) Edições Sílabo.

Lebkiri, N., Daoudi, M., Abidli, Z., Elturk, J., Soulaymani, A., Khatori, Y., el Madhi, Y., & Benattou, M. (2021). Using Machine Learning for Prediction Students Failure in Morocco: an Application of the CRISP-DM Methodology. *International Journal of Education and Information Technologies*, *15*, 344–352. https://doi.org/10.46300/9109.2021.15.36

Lee, S., Choi, K., & Yoo, D. (2020). Predicting the insolvency of smes using technological feasibility assessment information and data mining techniques. *Sustainability (Switzerland)*, *12*(23), 1–17. https://doi.org/10.3390/su12239790

Ma, B., Wang, Y., & Li, Z. (2022). Application of data mining in basketball statistics. *Applied Mathematics and Nonlinear Sciences*, *0*(0). https://doi.org/10.2478/amns.2021.2.00182

Major, M. J. (2017). Positivism and «alternative» accounting research. *Revista Contabilidade e Financas*, *28*(74), 173–178. https://doi.org/10.1590/1808-057x201790190

Malakauskas, A., & Lakstutiene, A. (2021). Financial distress prediction for small and medium enterprises using machine learning techniques. *Engineering Economics*, *32*(1), 4–14. https://doi.org/10.5755/j01.ee.32.1.27382

Mizik, N., & Hanssens, D. M. (2018). *Handbook of marketing analytics : methods and applications in marketing management, public policy, and litigation support*. Edward Elgar Pub.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464–473. https://doi.org/10.1016/j.dss.2011.10.007

Park, M., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of Machine Learning Models for Bankruptcy Prediction. *IEEE Access*, *9*, 124887–124899. https://doi.org/10.1109/ACCESS.2021.3110270

Park, S., Lee, C. W., Lee, S., & Lee, M. J. (2018). Landslide susceptibility mapping and comparison using decision tree models: A case study of Jumunjin Area, Korea. *Remote Sensing*, *10*(10). https://doi.org/10.3390/rs10101545

Peixoto, R. (2015). Pervasive Data Mining Engine. [Dissertação de mestrado, Universidade do Minho]. Repositório Universidade do Minho. https://hdl.handle.net/1822/40323

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, *36*(3), 1092–1113. https://doi.org/10.1016/j.ijforecast.2019.11.005

PORDATA - Estatísticas, gráficos e indicadores de Municípios, Portugal e Europa. https://www.pordata.pt/portugal/pequenas+e+medias+empresas+em+percentagem+do+total+de+empresas+total+e+por+dimensao-2859-248026 (acedido a 23/07/2023)

Portal Eportugal - https://eportugal.gov.pt/inicio/espaco-empresa/guia-a-a-z/cid-0-faseneg-2-falencia (acedido a 23/08/2022)

Putrada, A. G., Abdurohman, M., Perdana, D., & Nuha, H. H. (2022). Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey. *IEEE Access*, *10*, 45137–45178. https://doi.org/10.1109/ACCESS.2022.3169765

Ross Quinlan, by J., Kaufmann Publishers, M., & Salzberg, S. L. (1994). Programs for Machine Learning. *Machine Learning*, *16*, 235–240.

Shirata, C. Y., & Terano, T. (2000). Extracting Predictors of Corporate Bankruptcy: Empirical Study on Data Mining Methods. *Knowledge Discovery and Data Mining, Proceedings*, *1805*, 204–207. https://doi.org/https://doi.org/10.1007/3-540-45571-X_25

Simić, D., Kovačević, I., & Simić, S. (2012). Insolvency prediction for assessing corporate financial health. *Logic Journal of the IGPL*, *20*(3), 536–549. https://doi.org/10.1093/jigpal/jzr009

Stanisic, M., Stefanovic, D., Arezina, N., & Mizdrakovic, V. (2013). Analysis of Auditor's Reports and Bankruptcy Risk in Banking Sector in the Republic of Serbia. *Amfiteatru Economics*, *15*(34), 431–441. https://ssrn.com/abstract=2291496

Sun, J., & Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, *21*(1), 1–5. https://doi.org/10.1016/j.knosys.2006.11.003

Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*, *39*(5), 769–787. https://doi.org/10.1002/for.2661

Tian, J. X., & Zhang, J. (2022). Breast cancer diagnosis using feature extraction and boosted C5.0 decision tree algorithm with penalty factor. *Mathematical Biosciences and Engineering*, *19*(3), 2193–2205. https://doi.org/10.3934/MBE.2022102

Vanderlooy, S., & Hüllermeier, E. (2008). A critical analysis of variants of the AUC. *Machine Learning*, *72*(3), 247–262. https://doi.org/10.1007/s10994-008-5070-x

Wei-Yin Loh. (2008). Classification and Regression Tree Methods. *Encyclopedia of Statistics in Quality and Reliability*. https://doi.org/https://doi.org/10.1002/9780470061572.eqr492

Wilson, N., Wright, M., & Altanlar, A. (2014). The survival of newly-incorporated companies and founding director characteristics. *International Small Business Journal: Researching Entrepreneurship*, *32*(7), 733–758. https://doi.org/10.1177/0266242613476317

Yousaf, U., Jebran, K., & Wang, M. (2022). A comparison of static, dynamic and machine learning models in predicting the financial distress of chinese firms. *Romanian Journal of Economic Forecasting*, *25*(1), 122–138.

Zhou, L., Lu, D., & Fujita, H. (2015). The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches. *Knowledge-Based Systems*, *85*, 52–61. https://doi.org/10.1016/j.knosys.2015.04.017

Anexos

Anexo A – Medidas descritivas dos indicadores

Tabela A1: Descrição das variáveis demográficas

Variável	Descrição	Tipo	Unidade de medida	Descritivas
Anos_atividade	Idade da empresa à data do último ano de contas disponível	QD	Anos	RV: 707.291; M: 15,00; DP: 13,307
Unipessoal	Empresa unipessoal	QN	n.a.	RV: 707.291; Mo: Coletiva (71,60%)
Setor	Setor da empresa	QN	n.a.	RV: 707.291; Mo: Setor 7 (24,90%)
Região	Localização da empresa	QN	n.a.	RV: 707.291; Mo: AML (29,80%)
Dimensão	Dimensão da empresa	QO	n.a.	RV: 707.291; Mo: Pequena (80,30%)
FormaJuridica	Forma jurídica da empresa	QN	n.a.	RV: 707.291; Mo: LDA (94,10%)
Proporcao_Feminina	Proporção de mulheres no conselho de administração	QC	proporção	RV: 37.300; M: 0,24; DP: 0,31

Notas: DP: Desvio-padrão; M: Média; Mo: Moda; n.a.: não aplicável; QC: Quantitativa contínua; QD: Quantitativa discreta; QN: Qualitativa nominal; QO: Quantitativa ordinal; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A2: Descrição das variáveis de rentabilidade n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
ROAtivo	Rentabilidade do ativo tendo em conta o resultado operacional	QC	proporção	RV: 685.489; M: 0,00; DP: 0,37
RLAtivo	Rentabilidade do ativo	QC	proporção	RV: 685.515; M: -0,02; DP: 0,36
ROVendas	Rentabilidade operacional das vendas	QC	proporção	RV: 642.000; M: -0,04; DP: 0,63
RCP	Rentabilidade do capital próprio	QC	proporção	RV: 678.585; M: 0,13; DP: 1,13
MRADGFI	Margem de resultado antes de dep., gastos de fin. e impostos	QC	proporção	RV: 636.843; M: 0,11; DP: 0,26
ML	Margem de lucro	QC	proporção	RV: 631.888; M: 0,05; DP: 0,25

Notas: dep.: depreciações; DP: Desvio-padrão; fin.: financiamento; M: Média; QC: Quantitativa contínua; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A3: Descrição das variáveis de crescimento n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
Var.RCP	Variação RCP	QC	%	RV: 544.438; M: -0,50; DP: 12,37
Var.Empr.	Crescimento nº empregados	QC	%	RV: 601.031; M: 0,05; DP: 0,41
Var.ativo	Variação ativo total	QC	%	RV: 558.770; M: 0,27; DP: 1,15
Var.Vendas	Variação vendas	QC	%	RV: 523.218; M: 0,31; DP: 1,53

Notas: DP: Desvio-padrão; M: Média; nº: número; QC: Quantitativa contínua; RCP: Rentabilidade do capital próprio; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A4: Descrição das variáveis de endividamento n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
End	Endividamento	QC	proporção	RV: 693.042; M: 0,69; DP: 0,65
Solv	Solvabilidade	QC	proporção	RV: 680.776; M: 4,36; DP: 15,64
AF	Autonomia Financeira	QC	proporção	RV: 680.776; M: 0,30; DP: 0,65
PCAtivo	Passivo corrente/Ativo	QC	proporção	RV: 689.760; M: 0,44; DP: 0,48
FCDLP	Fluxo de caixa/Dívida de longo prazo	QC	proporção	RV: 317.233; M: 1,44; DP: 6,39
DLPAtivo	Dívida de longo prazo/Ativo	QC	proporção	RV: 472.753; M: 0,23; DP: 0,31
DLPCP	Dívida de longo prazo/Capital Próprio	QC	proporção	RV: 468.484; M: 0,98; DP: 4,97

Notas: DP: Desvio-padrão; M: Média; QC: Quantitativa contínua; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A5: Descrição das variáveis da estrutura do ativo n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
ACAtivo	Proporção do ativo corrente no total do ativo	QC	proporção	RV: 689.760; M: 0,72; DP: 0,29
FCAtivo	Fluxo de caixa/Ativo	QC	proporção	RV: 685.498; M: 0,02; DP: 0,36

Notas: DP: Desvio-padrão; M: Média; QC: Quantitativa contínua; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A6: Descrição das variáveis de liquidez n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
LG	Liquidez geral	QC	proporção	RV: 675.359; M: 10,07; DP: 33,61
ACPassivo	Ativo corrente/Passivo	QC	proporção	RV: 680.804; M: 3,82; DP: 11,20
FMAtivo	Fundo de maneio/Ativo	QC	proporção	RV: 641.149; M: 0,22; DP: 0,31
FCRO	Fluxo de caixa/Resultado operacional	QC	proporção	RV: 635.664; M: 0,08; DP: 0,24

Notas: DP: Desvio-padrão; M: Média; QC: Quantitativa contínua; RV: Respostas válidas

Fonte: Elaboração própria

Tabela A7: Descrição das variáveis de rotação n-1

Variável	Descrição	Tipo	Unidade de medida	Descritivas
VendasAtivo	Rotação do ativo	QC	proporção	RV: 686.461; M: 1,40; DP: 1,63
VendasAC	Rotação do ativo corrente	QC	proporção	RV: 685.779; M: 2,38; DP: 3,25

Notas: DP: Desvio-padrão; M: Média; QC: Quantitativa contínua; RV: Respostas válidas

Fonte: Elaboração própria

Anexo B – Fórmulas de cálculo das variáveis independentes

Tabela B1: Fórmula de cálculo das variáveis de rentabilidade

Variável	Fórmula de cálculo
ROAtivo	Resultado operacional/Total ativo
RLAtivo	Resultado líquido/Total ativo
ROVendas	Resultado operacional/Vendas
RCP	Resultado líquido/Total capital próprio
MRADGFI	Resultado antes de dep., gastos de fin. e impostos/Receita operacional
ML	Resultado antes de impostos/Receita operacional

Notas: dep.: depreciações; fin.: financiamento

Fonte: Elaboração própria

Tabela B2: Fórmula de cálculo das variáveis de crescimento de n-2 para n-1

Variável	Fórmula de cálculo
Var.RCP	$\frac{RCP\ n-1}{RCP\ n-2}$ - 1
Var.Empr.	$\frac{Nº\ empregados\ n-1}{Nº\ empregados\ n-2}$ - 1
Var.ativo	$\frac{Total\ ativo\ n-1}{Total\ ativo\ n-2}$ - 1
Var.Vendas	$\frac{Vendas\ n-1}{Vendas\ n-2}$ - 1

Notas: nº: número; RCP: Rentabilidade do capital próprio

Fonte: Elaboração própria

Tabela B3: Fórmula de cálculo das variáveis de endividamento

Variável	Fórmula de cálculo
End	Total passivo/Total ativo
Solv	Total capital próprio/Total passivo
AF	Total capital próprio/Total ativo
PCAtivo	Passivo corrente/Ativo
FCDLP	Fluxo de caixa/Dívida de longo prazo
DLPAtivo	Dívida de longo prazo/Total ativo
DLPCP	Dívida de longo prazo/Capital Próprio

Fonte: Elaboração própria

Tabela B4: Fórmula de cálculo das variáveis de estrutura do ativo

Variável	Fórmula de cálculo
ACAtivo	Ativo corrente/Total ativo
FCAtivo	Fluxo de caixa/Total ativo

Fonte: Elaboração própria

Tabela B5: Fórmula de cálculo das variáveis de liquidez

Variável	Fórmula de cálculo
LG	Ativo corrente/Passivo corrente
ACPassivo	Ativo corrente/Total passivo
FMAtivo	Fundo de maneio/Total ativo
FCRO	Fluxo de caixa/Resultado operacional

Notas: Fundo de maneio = Ativo corrente – Passivo corrente

Fonte: Elaboração própria

Tabela B6: Fórmula de cálculo das variáveis de rotação

Variável	Fórmula de cálculo
VendasAtivo	Vendas/Total ativo
VendasAC	Vendas/Ativo corrente

Fonte: Elaboração própria

Anexo C – Correlações entre as variáveis independentes

Tabela C1: Correlações entre as variáveis de crescimento para n-1

	Var.Empr._n1	Var.ativo_n1	Var.Vendas_n1	Var.RCP_n1
Var.Empr._n1	1,00			
Var.ativo_n1	0,13	1,00		
Var.Vendas_n1	0,13	0,39	1,00	
Var.RCP_n1	0,01	0,00	0,02	1,00

Fonte: Elaboração própria

Tabela C2: Correlações entre as variáveis de endividamento para n-1

	End_n1	Solv_n1	AF_n1	PCAtivo_n1	FCDLP_n1	DLPAtivo_n1	DLPCP_n1
End_n1	1,00						
Solv_n1	-0,27	1,00					
AF_n1	-0,97	0,27	1,00				
PCAtivo_n1	0,72	-0,23	-0,75	1,00			
FCDLP_n1	-0,16	0,17	0,16	-0,04	1,00		
DLPAtivo_n1	0,42	-0,16	-0,41	-0,08	-0,22	1,00	
DLPCP_n1	0,01	-0,04	-0,01	-0,07	-0,05	0,18	1,00

Fonte: Elaboração própria

Tabela C3: Correlações entre as variáveis da estrutura do ativo para n-1

	ACAtivo_n1	FCAtivo_n1
ACAtivo_n1	1,00	
FCAtivo_n1	-0,04	1,00

Fonte: Elaboração própria

Tabela C4: Correlações entre as variáveis de liquidez para n-1

	LG_n1	ACPassivo_n1	FMAtivo_n1	FCRO_n1
LG_n1	1,00			
ACPassivo_n1	0,59	1,00		
FMAtivo_n1	0,11	0,02	1,00	
FCRO_n1	0,08	0,12	-0,03	1,00

Fonte: Elaboração própria

Tabela C6: Correlações entre as variáveis de rentabilidade para n-1

	ROAtivo_n1	RLAtivo_n1	ROVendas_n1	RCP_n1	MRADGFI_n1	ML_n1
ROAtivo_n1	1,00					
RLAtivo_n1	0,99	1,00				
ROVendas_n1	0,49	0,48	1,00			
RCP_n1	0,05	0,04	0,07	1,00		
MRADGFI_n1	0,54	0,53	0,70	0,10	1,00	
ML_n1	0,59	0,57	0,80	0,12	0,89	1,00

Fonte: Elaboração própria

Tabela C7: Correlações entre as variáveis de rotação para n-1

	VendasAtivo_n1	VendasAC_n1
VendasAtivo_n1	1,00	
VendasAC_n1	0,70	1,00

Fonte: Elaboração própria

Anexo D – Avaliação dos *journals* pelo *Web of Science*

ID	Categoria do *journal*	Quartil *WoS 2021*
1	*Economics*	Q3
	Management	Q4
	Mathematics, Interdisciplinary Applications	Q3
2	*Economics*	Q4
3	*Economics*	Q2
4	*Economics*	Q3
5	*Computer Science, Information Systems*	Q2
	Engineering, Electrical & Electronic	Q2
	Telecommunications	Q2
6	*Economics*	Q3
7	*Environmental Sciences*	Q2
	Environmental Studies	Q2
	Green & Sustainable Science & Technology	Q3
8	*Economics*	Q2
	Management	Q4
9	*Economics*	Q1
	Management	Q2
10	*Business, Finance*	Q3
	Economics	Q3
	Mathematics, Interdisciplinary Applications	Q2
	Social Sciences, Mathematical Methods	Q3
11	*Computer Science, Artificial Intelligence*	Q1
	Computer Science, Interdisciplinary Applications	Q1
12	*Computer Science, Artificial Intelligence*	Q1
	Engineering, Electrical & Electronic	Q1
	Operations Research & Management Science	Q1
13	*Computer Science, Artificial Intelligence*	Q1
14	*Operations Research & Management Science*	Q1
15	*Computer Science, Artificial Intelligence*	Q1
	Computer Science, Information Systems	Q1
	Operations Research & Management Science	Q1
16	*Economics*	Q2
	Management	Q4
17	*Computer Science, Artificial Intelligence*	Q1
	Engineering, Electrical & Electronic	Q1
	Operations Research & Management Science	Q1
18	*Computer Science, Artificial Intelligence*	Q1
19	*Computer Science, Artificial Intelligence*	Q4
20	*Statistics & Probability*	Q1
21	*Computer Science, Artificial Intelligence*	Q4

Fonte: Elaboração própria

Anexo E – Comparação do modelo preditivo tendo em conta a variável da proporção feminina

		Modelos	
		Com proporção feminina	**Sem proporção feminina**
	Algoritmo	*CHAID*	*CHAID*
	Ensembles	-	-
Treino	**Especificidade**	76,17%	75,80%
	Sensibilidade	83,21%	82,90%
	Precisão	77,64%	77,40%
	F1-Score	0,80	0,80
	***Accuracy* (PCCC)**	79,68%	79,30%
	AUC	0,88	0,88
Teste	**Especificidade**	76,16%	76,00%
	Sensibilidade	82,94%	82,50%
	Precisão	25,95%	25,70%
	F1-Score	0,40	0,39
	***Accuracy* (PCCC)**	76,78%	76,60%
	AUC	0,88	0,88

Fonte: Elaboração própria