UMA APLICAÇÃO DE MACHINE LEARNING PARA PREVISÃO DE INSOLVÊNCIA: METODOLOGIA ADABOOST

José Augusto Gonçalves do Canto, joseaugustocanto@outlook.com, UTAD e UCAM - Universidade Candido Mendes
Amélia Cristina Ferreira da Silva, acfs@iscap.ipp.pt, CEOSP.PP - Centro de Estudos Organizacionais e Sociais do Politécnico do Porto
Gabriela Leite, gabriela_leite@sapo.pt, CEPESE - Centro de Estudos em População, Economia e Sociedade
Carlos Machado-Santos, cmsantos@utad.pt, UTAD e CEPESE - Centro de Estudos em População, Economia e Sociedade

RESUMO: Em economias de mercado livre, a insolvência de empresas é um fenómeno relativamente frequente, o que afeta inevitavelmente o grau de confiança dos investidores, podendo mesmo criar constrangimentos às transações. Assim, é importante que aqueles disponham de instrumentos capazes de antecipar as situações de insolvência e reduzir o risco financeiro das operações económicas. O objetivo deste artigo é desenvolver um modelo através da metodologia *Adaboost* para, com um ano de antecedência, prever a insolvência de pequenas e médias empresas. O modelo proposto, quando comparado com um modelo paramétrico tradicional, sugere um bom resultado. Para o efeito, utilizou-se a base de dados dos indicadores financeiros de 243 PMEs portuguesas do setor agroindustrial disponíveis no Sistema de Análise de Balanços Ibéricos. Os indicadores extraídos – liquidez de curto prazo e capacidade de gerar resultados adequados à dimensão – foram os indicadores mais relevantes estatisticamente tanto no modelo proposto, como no modelo de Regressão Logística.

PALAVRAS-CHAVE: Insolvência, *Bagging*, Árvore de Decisão, Treinamento Supervisionado, Matriz de Confusão.

ABSTRACT: Insolvency is a natural phenomenon for firms that operate in open market economies. The presence of potential insolvency makes difficult the economic transactions, which are based on trust. It is of crucial importance for economic agents the use of models that may predict and anticipate insolvency situations, reducing financial risks of economic operations. The aim of this study is to develop a model based on the *Adaboost methodology*, in order to predict the insolvency of Portuguese Small and Medium-sized Enterprises in the agro-industrial sector, with one year in advance. The database consists on financial indicators of 243 companies, available at Iberian Balance Analysis System. The propose model reveals a robust result when compared with traditional parametric models. The results show that two indicators – "short-term liquidity" and "capacity to generate results appropriate to the size" – were the most statistically relevant, both in the Proposed Model and the Logistic Regression model.

KEYWORDS: Insolvency, Bagging, Decision Tree, Supervised Training, Confusion Matrix.

1. INTRODUÇÃO

O natural ambiente de incerteza decorrente da dinâmica do mercado capitalista impulsiona o estudo de modelos matemáticos para o assunto previsão de insolvência das empresas, para desenvolver instrumentos que auxilie a tomada de decisão empresarial. Nestes estudos, conforme Breiman (2001) existem duas culturas no uso de modelos matemáticos: A primeira, tradicional na comunidade estatística, denominada de *data modeling culture* assume de maneira geral o modelo $r(x) = \beta 0 + \beta ixi$, o principal objetivo está na interpretação dos parâmetros βi's sujeita às hipóteses de normalidade, linearidade e homocedasticidade. A segunda, depende da evolução dos computadores, denominada de *algorithmic modeling culture* que domina a comunidade de *machine learning.*

Iniciado nos anos 30 a previsão de falência resumia-se a estudos comparativos de indicadores financeiros isolados. Nos anos 60 com a publicação de Altmam (1968) os estudos ganharam importante impulso ao relacionar os diversos indicadores financeiros com as técnicas estatísticas multivariadas. Nos anos 80, os modelos de análise discriminante dividiram espaço no estudo de previsão com os modelos logísticos, os quais apresentam seus resultados em forma de probabilidade acumulada, num avanço na qualidade interpretativa da previsão, ao substituírem os escores lineares dos modelos paramétricos. A evolução tecnológica nos anos 90 imprimiu alternativas no estudo de previsão de falência ao incorporar algoritmos de *machine learning* advindos da Inteligência Computacional, como por exemplo, Árvores de Decisão, Teoria de Redes Neurais, Teoria de Algoritmos Genéticos e da Teoria de Algoritmos Fuzzy.

Dentre as opções, Quinlan (1986) já destacava a maior facilidade de compreensão do algorítmo Árvores de Decisão por ser fortemente intuitivo, mas com problema de ter tendência de gerar modelos "superajustados" ou ("*overfitting"*) problema mais tarde confirmado por Kothari (2001). Isto acontece quando o modelo classifica bem o conjunto original de treinamento mas apresenta risco de degradar o seu desempenho para novos dados, por essa razão a técnica *tree bagging* de Breiman (1996) associa o processo de *bagging* a Árvores de Decisão para reduzir a instabilidade deste modelo.

Uma outra proposta para reduzir a instabilidade da Árvores de Decisão é a abordagem *boosting* de Schapire (1990), cuja principal diferença em relação ao processo *bagging* é a re-amostragem dos conjuntos de dados que gera aprendizados complementares, isto é, enquanto a metodologia *bagging* gera Árvores de Decisão paralelas a metodologia *boosting* gera réplicas de forma sequencial.

2. ENQUADRAMENTO TEÓRICO

Embora a utilização dos algoritmos de aprendizado de máquina nos trabalhos de finanças seja relativamente nova, diversos trabalhos que aprofundam o estudo de análise de risco de crédito ao utilizar os algoritmos para antecipar problemas financeiros estão sendo publicados (e.g., Auria & Moro, 2009; Brown, 2012; Butaru *et. al.*, 2016; Sealand, 2018). Nesta área, Dietterich (2000), Deng (2016), Bagherpour (2017), Addo *et. al.*, (2018) e Tokpavi (2018) comparam com bons resultados com o modelo estatístico tradicional Regressão Logística - procedimento seguido por este trabalho para previsão de insolvência das PMEs portuguesas do setor agroindustrial.

As réplicas Árvores de Decisão são algoritmos construídos por uma função conhecida como função-impureza. A função procura exaustivamente por meio de um processo recursivo sempre minimizar margem de erro, ela é mínima quando todos os dados pertencem à mesma classe e máxima quando os dados são linearmente distribuídos através das classes.

Segundo (Sutton, 2005) as funções-impurezas – Função Entropia e Gini Index – são citadas como as mais utilizadas para uma árvore de classificação (Eq. 1 e 2):

$$Entropia(N) = \sum_{j=1}^{m} -p_j \, log_2 p_j \quad [1]$$

$$Gini(N) = \sum_{j \neq m}^{m} -p_j \, p_{m=} 1 - \sum_{j=1}^{m} P_j{}^2 \quad [2]$$

Onde: *N* é o conjunto de exemplos; *m* é o conjunto de classes; p_j é a proporção de *N* pertencer à classe *j*, tendo então (Eq. 3):

$$p_j = \frac{|N_j|}{N} \quad [3]$$

O procedimento de crescimento da árvore tenta encontrar o caminho ótimo, pela seleção de atributos, uma das medidas conhecidas de seleção de atributos é o Ganho de Informação (Eq. 4).

$$\Delta Ganho(N, t) = Entropia(N) - Entropia(N_l) - Entropia(N_r) \quad [4]$$

Onde: t é o atributo corrente; $Entropia(N)$ é a impureza do nó corrente; $Entropia(N_l)$ é a impureza do nó esquerdo; $Entropia(N_r)$ é a impureza do nó direito; $\Delta Ganho(N, t)$ é o ganho do atributo *t* sobre o conjunto *N*.

A metodologia utiliza para treinamento supervisionado dos exemplos constituídos de indicadores financeiros para a construção da árvore. Assume-se o pressuposto de acumulação de informações nas demonstrações contábeis. Conforme Beaver (2006), o mesmo irá ocorrer com os indicadores financeiros o que justifica o seu uso como preditores ou estimadores da probabilidade de insolvência das empresas.

$$Prob\ (insolvência) = f(indicadores\ financeiros)$$

O conceito de insolvência aplicado para orientar o treinamento supervisionado está acordo com o n.º 2 do Artigo 3º do Código da Insolvência e da Recuperação de Empresas, descrito por Figueiredo (2018) "é considerado em situação de insolvência o devedor que se encontre impossibilitado de cumprir as suas obrigações vencidas, são também considerados insolventes quando o seu passivo seja manifestamente superior ao ativo, avaliados segundo as normas contabilísticas aplicáveis".

Na metodologia *bagging* cada réplica de árvore funciona como um classificador treinado, o conjunto de réplicas gera um comitê de árvores, que através do voto prevê um novo dado. Na metodologia *boosting*, os conjuntos de dados re-amostrados são construídos especificamente para gerar aprendizados complementares e a importância do voto é ponderado com base no desempenho de cada modelo, em vez da atribuição de mesmo peso para todos os votos.

A metodologia proposta se refere a um método específico de treinamento de um classificador aprimorado, que utiliza o algoritmo *AdaBoost,* "*Adaptive Boosting*" uma combinação das ideias de b*agging* e b*oosting* para melhorar o desempenho do algoritmo de aprendizado supervisionado Árvore de Decisão. No metodo réplicas de árvore funcionam como classificadores fracos que convergem a um classificador forte.

Conforme Schapire (1990), no *AdaBoost* os classificadores fracos, assim denominados por serem classificadores com desempenho um pouco melhor do que as suposições aleatórias são aprimorados sucessivamente através da atribuição de pesos (Eq. 5):

$$F_T(x) = \sum_{t=1}^{T} f_t(\mathrm{x}) \quad [5]$$

Cada f_t é um classificador fraco que gera uma hipótese $h_{(xi)}$ para cada amostra de conjunto de treinamento formado por *x* exemplos. Em cada iteração *t* um classificador fraco é selecionado e recebe um coeficiente $\propto_t$ para totalizar o resultado do erro $\in_t$ $(Eq. 6)$.

$$\in_t = \sum_i \in \left[F_{t-1}(x_i) + \propto_t h_{(xi)}\right] \quad [6]$$

$F_{t-1}(x_i)$ é um classificador fraco aprimorado na interação anterior e $\propto_t h_{(xi)}$ é o classificador fraco que será adicionado ao classificador final. Em cada iteração do processo de treinamento um peso $\omega_{i,t}$ é atribuído a cada amostra do conjunto de treinamento, esses pesos são usados para informar o treinamento do classificador fraco para priorizar as árvores com pesos altos.

No caso deste artigo de classificação binária será utilizado o algorítmo chamado de Adaboost.M1 para previsão da classe 0 ou 1 , que define o erro como: $\in_t = \Pr[\mathrm{ht(xi)} \neq \mathrm{yi}]$ e descarta o classificador fraco de hipótese $h_{(xi)}$ com erro superior a 0,5.

Input: sequência de N exemplos $((x_{1,}, y_{1,}) \ldots, (x_{n,}, y_n))$ com classes $y_i \in y\{1,2\}$; distribuição D sobre os exemplos N; número de interações inteiras T

Inicializar o vetor peso: $\omega_i^1 = D(i)\ para\ i = 1, \ldots, N$

Fazer para: $t = 1,2, \ldots, N$

1. Calcular o erro de $h_{(t)}$: $\in_t = \sum_{i=1}^{n} Pr[\mathrm{ht(xi)} \neq \mathrm{yi}]$
2. Se $\in_t > 0{,}5$ $fazer\ T = t - 1\ e\ abortar\ loop$
3. Fazer $\propto_t = {\in_t}/{(1 - \in_t)}$
4. Calcular novos pesos para o vetor peso

$$\omega_i^{t+1} = \omega_i^t \propto_t^{1-[\mathrm{ht(xi)} \neq \mathrm{yi}]}$$

Output das hipóteses (Eq. 7):

$$h_{(x)} = arq\ max \sum_{t=1}^{T} (log\ {1}/{\propto_t})[\mathrm{ht(xi)} = \mathrm{yi}] \quad [7]$$

A técnica de seleção binária que servirá como referência estatística tradicional para validar a metodologia proposta utiliza indicadores contábeis logisticamente distribuídos, na forma de probabilidade acumulada entre os valores 0 e 1. Por apresentar a forma de probabilidade, fornece melhor qualidade interpretativa para a previsão, atributo significante na tomada de decisões. A distribuição logística descrita por Zavgren (1985) é um tipo especial de função sendo identificada mais precisamente como função de distribuição acumulada logística (Eq. 8).

$$Pi = E(Y = 1/\ Xi) = (e^{B0+B1x})/(1 + e^{B0+B1x}) \quad [8]$$

Onde: $\beta_0 + \beta_1 x_1$ são coeficientes lineares sujeitos às hipótese estatísticas.

Um dos primeiros trabalhos relevantes de análise logística é de Ohlson (1980) que utilizou oito indicadores financeiros e foi capaz de identificar com um ano de antecedência a falência de empresas com 89% de precisão. Por seu turno, Shumway (2001) e Hensher *et al*. (2007) também

utilizaram indicadores financeiros e a técnica logística para antecipar falência com bons resultados, 88% e 92%respetivamente.

3. METODOLOGIA

O objetivo é construir uma metodologia de *Machine Learning* para a previsão de insolvência das PMEs portuguesas do setor agronegócio através do algorítmo ensemble denominado *Adaboost*. A validação do modelo proposto segue a metodologia de trabalhos correlacionados ao utilizar um modelo estatístico tradicional como parâmetro de desempenho.

Os experimentos realizados neste trabalho podem ser divididos em dois grupos: ajustes e testes com modelagem logística e o modelo proposto, os experimentos são realizados separadamente tendo em comum somente as fases de definição dos dados.

A descrição metodológica resumida na Figura 1 somente inclui a metodologia experimental omitindo-se as etapas comuns a qualquer trabalho de pesquisa tais como, revisão bibliográfica e conclusão. Dividida em cinco fases, a saber: (1) Descrição dos dados (ou seja, dos indicadores), (2) limpeza dos dados, (3) seleção das variáveis, (4) ajuste (ou treinamento) e (5) testes.

Na etapa Descrição dos Dados, são descritos os indicadores que constituem potenciais variáveis de entrada para os modelos de previsão. Nas etapas Limpeza dos Dados, Seleção de Variáveis e Testes são comentadas as estratégias aqui utilizadas. Os detalhes experimentais são comentados na etapa Experimentos.

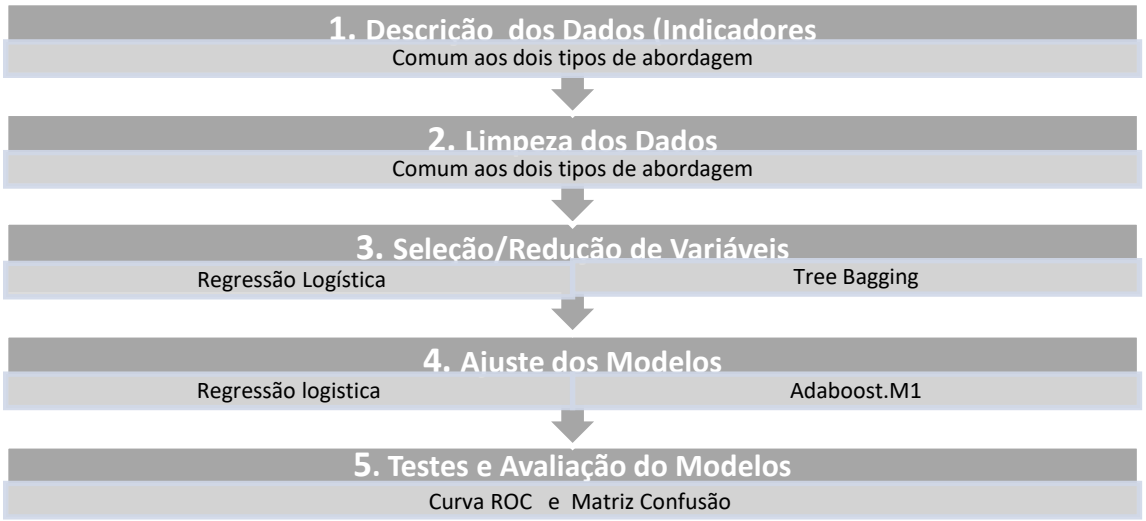


Figura 1: Metodologia experimental
Fonte: Elaboração própria

A base de dados utilizada contém indicadores financeiros europeus contidos no banco de dados da ferramenta de pesquisa SABI (Sistema de Análise de Balanços Ibéricos), a base inicial constou de 2.236 PMEs portuguesas do setor agroindustrial: agricultura, produção animal, caça e atividades dos serviços relacionados a silvicultura, exploração florestal, indústrias alimentares, bebidas, tabaco; couro e cortiça.

Foi adotado o conceito europeu de PME, publicado pelo Jornal Oficial da União Europeia de 20.5.2003, “A categoria das micro, pequenas e médias empresas (PME) é constituída por empresas

que empregam menos de 250 pessoas e cujo volume de negócios anual não excede 50 milhões de euros ou cujo balanço total anual não excede 43 milhões de euros “.

A totalidade das PMEs organizadas em “*cross-section”* observou o intervalo temporal 2007-2017 das publicações anuais dos respetivos indicadores financeiros das empresas contidas no banco de dados.

A partir da base inicial, foram adotados critérios para selecionar a amostra final, o primeiro critério a extração da base apenas as PMEs com indicadores financeiros completos na série. As empresas foram divididas em duas classes, empresas solventes e empresas insolventes, de acordo com os seguintes critérios de seleção:

- Empresa insolvente - Empresa com publicação um ano antes do Capital Próprio (CP) tornar-se negativo, em uma série de pelo menos três anos consecutivos negativos e, empresa com publicação um ano antes de ter abandonado a base por *default*.
- Empresa solvente: não contemplar capital próprio negativo no período em análise, 2007-2017.

Os critérios adotados para a escolha dos indicadores contemplam a integridade dos dados em relação a implantação do Sistema de Normalização Contabilística em 1 de Janeiro de 2010, as empresas solventes foram todas coletadas em 2017 e as empresas insolventes após 2010 devido o critério de três balanços seguidos com capital próprio negativo. Após coletada a amostra, foram selecionados 11 indicadores financeiros tradicionais, conforme Tabela 1.

Tabela 1: Indicadores utilizados

Literal	**Fórmula**
Rácio de liquidez corrente	Ativo circulante / Passivo líquido
Rácio de liquidez	(Ativo circulante - Estoques) / Passivo líquido
Rácio de liquidez dos acionistas	Capital próprio / Passivos fixos
Rácio de solvabilidade	(Capital próprio / Ativos totais)* 100
Alavancagem	((Passivos fixos + Dívidas financeiras) / Capital próprio) * 100
Margem de lucro	(Lucro Antes dos Impostos / Resultado Operacional) * 100
Rácio de liquidez dos accionistas	(Lucro Antes de Impostos / Capital próprio) * 100
Return on Capital Employed	(Resultados antes de despesas fiscais + financeiras e despesas similares) / (Capital Próprio + Passivos fixos)) * 100
Return on Total Assets	(Resultados antes do imposto / Total ativo) * 100
Capacidade de cobrir juros	Exploração de Resultados / Despesas financeiras e despesas similares
Stock Turnover	Resultado operacional / Stock

Fonte: Elaboração própria

A limpeza dos dados é um tratamento realizado sobre os dados selecionados, de forma a assegurar a qualidade (completude, veracidade e integridade) dos factos por eles representados. São tarefas comuns da limpeza de dados: preencher valores faltantes; identificar *outliers* e suavizar ruídos e corrigir informações errôneas ou inconsistentes. Além de dados faltantes são identificados *outliers* que, mostraram-se inconsistentes com a realidade. Desta forma, neste trabalho foram necessários ajustes nos dados referentes às duas primeiras tarefas.

A seleção das variáveis preditoras serão efetuadas separadamente, porém algumas considerações comuns serão efetuadas antes, como a verificação de existência de alta correlação entre as variáveis preditoras. Para selecionar as variáveis da modelagem Regressão Logística é aplicado o teste paramétrico Wald e verificada a hipótese nula ao nível de 5%. Para a modelagem *Tree-Bagging* é verificada a importância dos atributos medida pelo erro de classificação das observações "*out-of-bag",* o processo compreende na retirada sucessiva de variáveis preditoras para verificar a variação do erro de classificação com a falta. Conforme Arlot *et al.* (2010), para encontrar o melhor conjunto de preditores é testado o erro de validação cruzada 10-fold e seleciona-se o conjunto de indicadores com o menor erro, no processo os exemplos são divididos em dez partes "*folds"*, nove utilizadas para treinamento e uma para teste de maneira circular e sucessiva.

Os modelos são ajustados separadamente, na Regressão Logística são verificados os valores dos coeficientes gerados para a montagem da função *logit* preditora de insolvência empresarial. Na

metodologia *bagging* são geradas 200 Árvores de Decisão e verificado e erro de classificação, é esperado a redução do erro em função dos números de cópias *bootstrap*. As 200 árvores formam o comitê de votos na qual cada cópia *bootstrap* tem um voto para a previsão de insolvência da PME, com isto a metodologia enfrenta o problema de *overfitting* do modelo Árvore de Decisão.

Após a fase de ajustes, os modelos são testados e avaliados através de testes estatísticos. Os modelos são avaliados pela quantidade de acertos e tipos de erros, ao tentar diferenciar as empresas solventes de empresas insolventes pode acontecer dois tipos de erros: Erro do tipo I, relacionado a um resultado de insolvência quando a empresa é solvente e erro do tipo II que representa a possibilidade de selecionar a empresa como solvente quando é insolvente. Para verificar os acertos e os erros a metodologia *Machine Learnig* empresta um método da área médica utilizado para avaliar a qualidade dos exames de saúde, que utiliza a tabela Matriz Confusão para contabilizar os resultados e a ferramenta Curva ROC que possibilita avaliação dos exames para diversos pontos de cortes.

A ferramenta Matriz de Confusão oferece o número de classificações corretas e incorretas *versus* as classificações preditas para cada classe em um conjunto de exemplos dicotômicos, informações utilizadas para calcular as medidas de acurácia, sensibilidade e a especificidade, medidas efetivas de desempenho que relacionam os riscos de cometer Erro do tipo I ou Erro do tipo II. A sensibilidade está relacionada ao Erro do Tipo I, que é o risco de rejeitar H0. A especificidade está relacionada ao erro do Tipo II, não rejeitar H0. Curva ROC representa a sensibilidade e a especificidade para todos os possíveis valores de pontos de corte sob a curva, será utilizada para avaliar globalmente as metodologias utilizadas neste trabalho.
A Matriz de Confusão, exemplificada na tabela 2, contabiliza os dados necessários para os cálculos das métricas denominadas de acurácia, especificidade e sensibilidade. O resultado FP está relacionado ao Erro Tipo I e o FN está relacionado ao Erro Tipo II, VP - Verdadeiro positivo; VN - Verdadeiro negativo FP - Falso positivo; FN – Falso negativo.

Tabela 2: Modelo da Matriz Confusão

Insolvente previsto	VP	FP
Solvente previsto	FN	VN
Classes	Insolvente	Solvente

Fonte: Elaboração própria

Acurácia: Mede a probabilidade do resultado do teste estar classificado corretamente dado os exemplos totais: $(VP + VN)/T$. Sensibilidade: Corresponde à probabilidade do teste classificar corretamente uma empresa insolvente: $VP/(VP + FN)$. Especificidade: Corresponde à probabilidade do teste classificar corretamente uma empresa solvente $VN/(FP + VN)$.

4. RESULTADOS

Na base de dados inicial de 2.236 PMEs portuguesas do setor agroindustrial foram identificadas 2.058 empresas solventes e 178 insolventes, uma amostra claramente desbalanceada. Drummond *et al.* (2003) explica que acurácia e a capacidade de generalização dos modelos para problema de seleção sofrem influência do tamanho da amostra, do número de atributos e do balanceamento dos dados, tal fato, implica restrições na seleção.

Priorizado o problema do desbalanceamento dos dados, a solução adotada foi equilibrar a amostra para 356 empresas, que após o processo de limpeza dos dados, *outliers* e dados faltantes ficou reduzida à 243 empresas, 122 solventes e 121 insolventes.

Para adequar a complexidade dos modelos ao tamanho e a qualidade da amostra disponível, o processo de seleção dos atributos foi separado por metodologia, quando os atributos iniciais foram reduzidos para os mais significativos e os mais importantes. Todos os experimentos foram realizados utilizando-se a plataforma computacional Matlab® da Mathworks.

4.1 Seleção das variáveis de entrada

Para efeitos de síntese e simplicidade, as variáveis, ou atributos, assumem números de entrada, a Tabela 3 apresenta correspondência numérica dos atributos.

Tabela 3: Correspondência numérica dos atributos

1	Retorno sobre capital próprio
2	Retorno sobre capital investido
3	Retorno sobre o total do activo
4	Margem de lucro
5	Capacidade de cobrir juros
6	Stock Turnover
7	Rácio de liquidez corrente
8	Rácio de liquidez
9	Rácio de liquidez dos acionistas
10	Rácio de solvabilidade
11	Alavancagem

Fonte: Elaboração própria

Antes de ter sido aplicada as metodologias especificas para selecionar as variáveis de entrada, foi verificado através da matriz de correlação descrita na tabela 4 as variáveis explicativas com alta correlação. Para o limiar de 0,5, verifica-se que os atributos (1 e 2), (1 e 3) , (1 e 4) , (1 e 11), são fortemente correlacionados, não é recomendável que estejam juntas na seleção de variáveis. Pelo mesmo motivo , as variáveis (2 e 3) , (2 e 4), (3 e 4), (3 e 10), (7 e 8) e finalmente (8 e 10).

Tabela 4 – Resultado da Matriz de Correlação

	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***
1	1										
2	0,622	1,000									
3	0,720	0,692	1,000								
4	0,610	0,524	0,781	1,000							
5	0,215	0,182	0,225	0,127	1,000						
6	0,079	0,100	0,239	0,104	0,048	1,000					
7	0,133	0,117	0,182	0,146	0,028	-0,031	1,000				
8	0,150	0,136	0,301	0,196	0,122	0,103	0,778	1,000			
9	0,074	0,012	0,148	0,074	0,143	0,071	0,115	0,141	1,000		
10	0,386	0,240	0,504	0,419	0,062	0,142	0,425	0,518	0,287	1,000	
11	-0,562	-0,169	-0,340	0,343	-0,023	-0,082	-0,124	-0,155	-0,114	-0,471	1,000

Fonte: Adaptado da plataforma Matlab

Além de ter verificado o nível de correlação entre as variáveis de entrada, verificou-se também a possibilidade relação espúria entre as variáveis de entrada e saída. No estudo, a variável de saída utilizada para o processo de treinamento supervisionado é uma variável dicotômica, com relação direta da situação líquida do capital próprio das respetivas PMEs, assume o valor um (1) para solvente e o valor zero (0) para insolvente.

Para evitar relações artificiais de causa e efeitos entre as variáveis de entrada e saída, as variáveis de entrada 1,2,9,10 e 11 não foram utilizadas no processo de aprendizado supervisionado por conterem o atributo capital próprio nas suas construções.

No teste Wald para a regressão logística a estatística p-valor é obtida por comparação entre a estimativa de máxima verossimilhança do parâmetro ($\widehat{\beta_J}$) e a estimativa de seu erro padrão, a razão resultante sob a hipótese $H_0: \beta_j = 0$ tem distribuição normal padrão (Eq. 9).

$$W_j = {\widehat{\beta_J}}\Big/{DP\widehat{\beta_J}} \quad [9]$$

O p-valor é definido como $P(\lfloor Z \rfloor > W_j)$, sendo que Z expressa a variável aleatória da distribuição normal padrão.

Utilizado o teste Wald, para selecionar do conjunto de 6 atributos os mais significantes, verifica-se na tabela 3 que as variáveis 3 e 8, ao nível de significância de 5%, rejeitam a hipótese nula (Retorno sobre o total do ativo e Rácio de liquidez). A tabela é descrita : Primeira coluna – variáveis estimadoras ; β_j – constantes correspondente a cada variável estimadora; $DP\widehat{\beta_J}$ — erro padrão dos coeficientes;Wald - para cada coeficiente para testar a hipótese nula, que corresponde a coeficiente zero contra a hipótese alternativa diferente de zero; pValue — *p*-value para *F*-statistic do teste de hipótese que corresponde o coeficiente igual a zero ou não. Se o valor do for maior do 0,05 a variável não é significativa ao nível de significância de 5% dado outras variáveis do modelo.

Tabela 5 – Estimação dos Coeficientes

Variáveis estimadoras	β_j	$DP\widehat{\beta_J}$	**Wald**	**pValue**
Intercepto	0.6641	0.3828	1.7348	0.0827
3	-0.3693	0.0580	-6.3659	1.9417e-10
4	-0.0313	0.0438	-0.7151	0.4745
5	0.0013	0.0011	1.1633	0.2446
6	0.0022	0.0023	0.9437	0.3453
7	0.2214	0.3023	0.7323	0.4639
8	-1.2665	0.5527	-2.2902	0.0220

Fonte: Resultado adaptado da plataforma Matlab

Para se estimar a importância dos atributos quando se utiliza a metodologia *tree bagging* para a seleção de variáveis é preciso inspecionar-se como o erro do conjunto varia com a acumulação das árvores. A importância dos estimadores pode ser observada através da permutação randômica dos dados *out-of-bag* pela retirada do estimador e verificado o incremento do erro devido a sua falta. O maior incremento de erro significa maior importância do estimador.

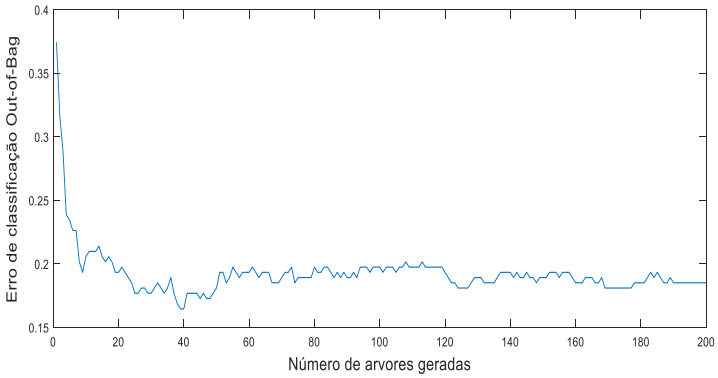
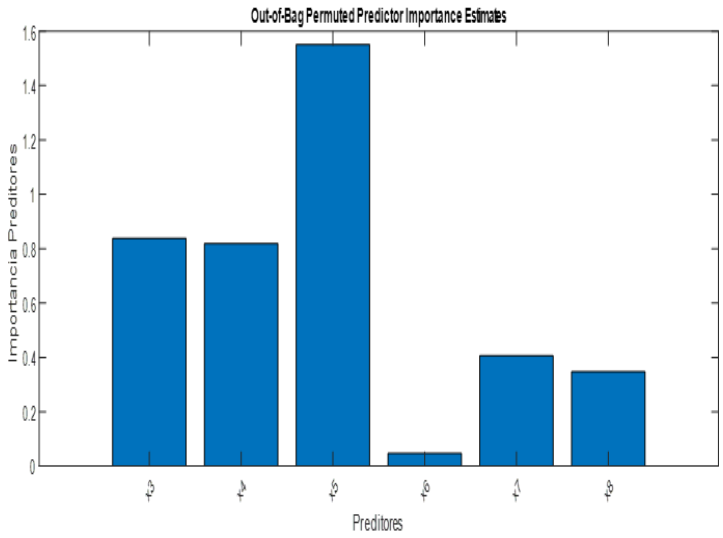


Figura 2: Variação do erro *out-of-bag* com o número de árvores geradas
Fonte : Resultado impresso da plataforma Matlab

Inicialmente, verifica-se como o erro das observações varia com o aumento das árvores do conjunto. É esperado que esse erro fique reduzido com o número de árvores. A Figura 2 apresenta o gráfico da variação deste erro com o número de árvores. Foram geradas 200 árvores e o gráfico mostra claramente o erro diminuindo, o que significa que o processo de *tree bagging* está adequado.

Para problemas de classificação como o deste trabalho é recomendável que o tamanho mínimo dos nós terminais seja um. Além disso, seleciona-se a raiz quadrada do número total de atributos para cada divisão de decisão nos nós, aleatoriamente.

A Figura 3 demonstra a importância dos atributos medida pelo erro de classificação das observações *out-of-bag*. O aumento do erro de classificação, devido à permutação dos dados, representa o quão importante é o atributo .



.

Figura 3: Importância do atributo, medida como o erro de classificação *out-of-bag*
Fonte: Extraido da plataforma Matlab

Pela ordem sugerida pelo método *tree bagging* a importância das seis variáveis mais importantes é 5, 3, 4, 7, 8 e 6. Contudo, o atributo 7 tem forte correlação com o atributo 8. Portanto, o atributo 7 foi excluído da lista. Selecionados os cincos atributos foi repetido o procedimento, o resultado na Figura 4 confirmou a seleção anterior.

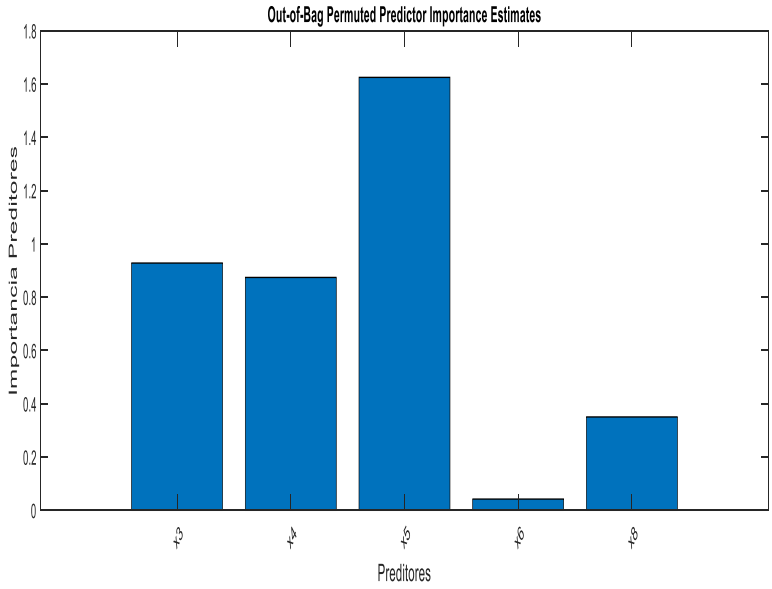


Figura 4: Importância dos atributos entre os 5 atributos selecionados
Fonte : Extraído da plataforma Matlab

A partir do conjunto de cinco variáveis foi selecionado outro conjunto de variáveis. A variável 6 foi descartada por ter importância muita distinta, o passo seguinte foi testar quatro combinações possíveis com as variáveis restantes.

As combinações geraram modelos de três variáveis ilustradas na Tabela 6 cuja representação mostra o erro de validação cruzada 10-fold como critério de seleção de variáveis.

Tabela 6: Combinações de três atributos testadas.

Combinação de atributos	Erro de validação cruzada
{3,4,5}	0.1975
{3,4,8}	0.2016
{3,5,8}	0.1893
{8,5,4}	0.1934

Fonte: Resultados extraídos da plataforma Matlab

Diante dos resultados obtidos, as variáveis selecionadas são as 3, 5, 8 (Retorno sobre o total do ativo, Rácio de liquidez , Capacidade de cobrir juros) por apresentar o menor erro de validação.

4.2 Ajustes e Avaliação dos resultados

Como resultado do juste do modelo Regressão logística a equação preditora pode ser descrita - a probabilidade de insolvência de uma PME com um ano antecedente :

$$\boldsymbol{P(Y=1) = {1}/{(1 - e^{-g(x)})}} \quad [10]$$

Onde

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots \beta_j x_j \ ;$$

Resultado do ajuste da Regressão Logística:

$$g(x) = \beta_0 + \ 0{,}664 - 0{,}3693.\,\text{Retorno sobre o total do activo} - \ 0{,}2665.\,\text{Racio de liquidez}$$

Na metodologia *Adaboost* a base do sistema proposto é uma Árvore de Decisão, cuja aprendizagem supervisionada utilizou como entrada o conjunto de três indicadores mais importantes, x3 = Retorno sobre o total do ativo, x8 = Rácio de liquidez e x5 = Capacidade de cobrir juros, como saída para o processo de treinamento foi adotado os valores de saída 0 e 1, que representam as classes de insolvência e solvência.

Os resultados foram avaliados através das métricas de acurácia, sensibilidade e especificidade calculadas com base nos dados apresentadas nas Matrizes de Confusão e da métrica AUC da curva ROC. Todos os dados foram extraídos dos modelos ajustados na plataforma Matlab. As tabelas 7 e 8 apresentam os resultados dos ajustes das metodologias Regressão Logística e *Adaboost*.

Tabela 7: Resultados do ajuste da metodologia Regressão Linear

Matriz Confusão			Métricas para avaliação	
Insolvente	104	20	Acurácia -	104+89/243=79,42%
Solvente	30	89	Sensibilidade -	104/(104+30)=77,61%
	0 Classe	1 Predita	Especificidade	89/(89 +20) =81,65%

Fonte: Elaboração própria

Tabela 8: Resultados do ajuste da metodologia *Adaboost*

Matriz Confusão			**Métricas para avalição**	
Insolvente	90	29	Acurácia -	105+90/243=80,25%
Solvente	19	105	Sensibilidade -	90/(90+19)=82,57%
	0 Classe	1 Predita	Especificidade -	105/(105+29)=78.36%

Fonte: Elaboração própria

As métricas apresentados na tabela 9 sugerem superioridade do modelo *Adaboost* em relação ao modelo tradicional de seleção Regressão Logística. O teste de Acurácia apresentou 80,25 % de probabilidade de acertar a previsão do estado de insolvência das PMEs portuguesas do setor agroindustrial com um ano de antecedência, enquanto o modelo tradicional 79,42 % de probabilidade. O teste de Sensibilidade do modelo proposto apresentou 82,57% de probabilidade de prever insolvência sendo a PME insolvente, o modelo tradicional 77,61 %. O teste de Especificidade do modelo proposto apresentou a probabilidade de 78,3635% de prever a solvência sendo a PME solvente, o modelo tradicional apresentou 81,65%. Porém, para o objetivo do modelo de prever insolvência o resultado da Sensibilidade é mais relevante.

O resultado 0,90 do teste de ajuste da medida AUC da curva ROC da metodologia proposta e 0,89 do tradicional indica qualidade superior do ajuste da metodologia proposta para prever insolvência das PME quando o ponto de corte das medidas de sensibilidade e especificidade são alterados.

Tabela 9: Resultados consolidados dos ajustes

	Acurácia	**Sensibilidade**	**Especificidade**	**AUC**
Regressão Logística	79,42 %	77,61%	81,65 %	0,89
Adaboost	80,25 %	82,57 %	78,36 %	0,90

Fonte: Elaboração própria

5. CONCLUSÃO

No objetivo de desenvolver através do algoritmo *Adaboost* um modelo da *algorithmic modeling culture* para predizer insolvência com um ano de antecedência o estudo de Beaver (2006) acompanhou Edmister (1972) na previsão de falência de pequenas empresas "através de certas razões financeiras e fazendo uso da técnica de análise discriminante pode-se predizer com certa antecipação e algum grau de confiabilidade a falência de uma pequena empresa".

Para o treinamento supervisionado foi utilizado como exemplo uma amostra equilibrada de razões financeiras de 243 PMEs portuguesas, empresas que empregam menos de 250 pessoas e cujo volume de negócios anual não excede 50 milhões de euros. A dificuldade encontrada foi a reduzida quantidade de registos de PMEs insolventes. Apesar da base inicial conter 2.236 PMEs do setor agroindustrial, a quantidade de empresas enquadradas no critério de insolventes eram apenas de

178, que após o processo de limpeza resultou apenas 121 empresas, restando duas opções para equilibrar a amostra: ou elevar a quantidade de insolventes através de métodos artificiais ou selecionar randomicamente 122 empresas solventes – tendo-se decidido pela segunda opção.

Para validar a metodologia proposta foram efetuadas comparações de métricas de desempenho com um modelo tradicional da *data modeling culture* a Regressão Logística. Na construção dos modelos o processo de seleção dos indicadores mais relevantes chamou atenção, apesar das metodologias diferentes os indicadores selecionados como mais relevantes foram iguais, liquidez de curto prazo e capacidade de geração de resultados adequados ao tamanho da empresa.

O resultado do confronto entre as medidas de desempenho do modelo proposto e o modelo tradicional sugere validação da metodologia *Adaboost*. A medida de desempenho considerada mais importante para validação foi a medida de Sensibilidade que quantifica a probabilidade de prever com antecedência a insolvência das PME realmente insolventes. O modelo proposto apresentou resultado de 82,57% e a Regressão Logística 77,61 % de probabilidade de prever insolvência de uma PME portuguesa.

O resultado sugere também a importância no aprofundamento de estudos que utilizam a metodologia *Adaboost* para o melhor entendimento do fenómeno insolvência para as PMEs portuguesas do setor agroindustrial.

REFERÊNCIAS

Addo, P. M., Guegan, D. & Hassani, B. (2018). *Credit Risk Analysis Using Machine and Deep Learning Models*. *SSRN*. Recuperado de https://doi.org/10.2139/ssrn.3155047

Auria, L., & Moro, R. A. (2009). Support Vector Machines (SVM) as a Technique for Solvency Analysis. SSRN. Recuperado de https://doi.org/10.2139/ssrn.1424949

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

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, Vol. 4, 40-79.

Beaver, W.H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research 4 (supplement), 71-111.

Bolarinwa, A. (2017). Machine learning applications in mortgage default prediction. University of Tampere. Recuperado de http://urn.fi/URN:NBN:fi:uta-201712122923

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140https://doi.org/10.1007/BF00058655

Brown, D. R. (2012). A Comparative Analysis of Machine Learning Techniques For Foreclosure Prediction. Nova Southeastern University. Recuperado de https://nsuworks.nova.edu/gscis_etd/105/

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W. & Siddique, A. (2016). Risk and risk management in the credit card industry. Journal of Banking & Finance. Recuperado de https://doi.org/10.1016/j.jbankfin.2016.07.015

Deng, G. (2016). Analyzing the Risk of Mortgage Default. University of California. Recuperado de https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Grace_Deng_thesis.pdf

Dietterich, T. (2000). An empirical comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Machine Learning, 40(2): 139-157.

Drummond, C.; & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: why undersampling beats over-sampling. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets.

Edmister, R.O. (1972). An empirical test of financial ratio: analysis for small business failure prediction. Journal of Financial and Quantitative Analysis, 7, 1477-1493

Figueiredo, H.M. (2018). 'O problema da recuperação de empresas em Portugal : Analise Crítica', Dissertação Mestrado, Instituto Superior de Contabilidade e Administração de Coimbra. Recuperado de https://comum.rcaap.pt/bitstream/10400.26/23121/1/Helena_Figueiredo.pdf

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the data boost-IM approach. SIGKDD Explorations, 6(1).

He, Y.,& Kamath, R. (2005). Bankruptcy prediction of small firms: in individual industries with the help of mixed industry models. Asia-Pacific Journal of Accounting & Economics, 12 (1), 19-36.

Hensher, D.A., & Stwart, J. (2007). Forecasting corporate bankruptcy: optimizing the performance of the mixed logit model. Abacus, Vol. 43, 3, 241-364.

Hsiao, S., & Whang, T. (2009). A study of financial insolvency prediction model for life insurers. Expert Systems with Applications, Vol.36, 3, 6100-6107.

Jain, A., & Zongker, D. (1997). Transactions on pattern analysis and machine intelligence. Expert Systems with Applications, Volume: 19, Issue: 2, 153 – 158.

Kothari, R., & Dong, M. (2001). Decision trees for classification: a review and some new results. Lecture Notes in Pattern Recognition, World Scientific Publishing. p. 241-252.

Kumar, P.R., & Ravi,V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research, Vol. 180, 1, 1-28

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

Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, p. 81-106

Sealand, J. C. (2018). Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models. Slippery Rock University. Recuperado de https://www.researchgate.net/profile/Jesse_Sealand/publication/326518013

Schapire, R. E.(1990) The strength of weak learnability. Mach Learn 5 (2), 197–227. Recuperado de https://dx.doi.org/10.1007/BF00116037

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. Journal of Business, 74, 101-124.

Sutton, C.D. (2005). Classification and Regression Trees,Bagging, and Boosting. Elsevier B.V. Handbook of statistics, Vol. 24, ISSN: 0169-7161

Tokpavi, H. S. H. C. S. (2018). Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. Recuperado de https://www.researchgate.net/publication/318661593

Zavgren, C.V. (1985). Assessing the vulnerability of failure of American industrial firms: a logistic analysis. Journal of Business.Vol 1, 19-45.