Previsão de insolvência das PME portuguesas do setor agroindustrial: Metodologia Tree-Bagging

José Augusto Canto
joseaugustocanto@outlook.com
Ucam-Universidade Candido Mendes

Amélia 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.

C. Machado-Santos
cmsantos@utad.pt
UTAD e CEPESE Centro de Estudos em População, Economia e Sociedade

A insolvência é um fenómeno natural das empresas que operam em economias de mercado livre. No entanto, a ameaça da insolvência dificulta as transações económicas baseadas na confiança. Neste contexto, é importante que os agentes económicos 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 *tree-bagging* para prever a insolvência das pequenas e médias empresas (PME) portuguesas do setor agroindustrial, com um ano de antecedência. A base de dados utilizada é composta de indicadores financeiros de 243 empresas, disponíveis no Sistema de Análise de Balanços Ibéricos, todas do setor agroindustrial. O modelo proposto, quando comparado com um modelo paramétrico tradicional, sugere um bom resultado. Os indicadores extraídos - liquidez de curto prazo e capacidade de gerar resultados adequados ao porte - foram os indicadores mais relevantes estatisticamente tanto no modelo proposto, como no modelo de Regressão Logística.

Palavras chave: Insolvência; Bagging, Árvore de Decisão, Overfitting

1. INTRODUÇÂO

Ao longo dos anos diversos estudos têm utilizado variadas técnicas para desenvolver modelos de previsão de insolvência, conforme (Hsiao et al., 2009) a utilização de algoritmos matemáticos para a solução destes problemas acompanha a evolução dos computadores, dentre os algoritmos presencia-se o crescimento do interesse da metodologia "*machine learning*". Contribui para isso a necessidade natural de reduzir o grau de incerteza no ambiente corporativo.

Neste sentido a utilização deste instrumento tecnológico na academia na área financeira procura melhorar a compreensão do fenômeno insolvência das empresas. 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.

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

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 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.

A metodologia proposta utiliza a técnica *tree bagging* para treinamento supervisionado dos exemplos constituídos de indicadores financeiros. 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".

A técnica *tree bagging* explicada por (He et al., 2005) e (Guoh et al., 2004) é um classificador gerado por réplicas de Arvores de Decisão, que 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.

$$Entropia(N) = \sum_{j=1}^{m} -p_j\ log_2 p_j\ \ ;$$

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

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: $p_j = \frac{|N_j|}{N}$

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

$$\Delta Ganho(N,t) = Entropia(N) - Entropia(N_l) - Entropia(N_r)$$

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.

Para cada réplica é gerada uma Árvore de Decisão que funciona como um classificador treinado, o conjunto de réplicas gera um comitê de arvores, que através do voto prevê um novo dado, é razoável supor que essa predição é bem mais robusta que a predição de uma única árvore.

Para gerar múltiplas versões de Árvores de Decisão o método *Bagging* constrói amostras *bootstrap* a partir do conjunto de dados originais. Conforme (Breimam,1996), um conjunto de treinamento $\mathcal{L}$ consiste dos dados $\{(x_i\ ,\ y_i)\ , i\ =\ 1;:::;N\}$, onde : N é a quantidade de exemplos; x_i atributos ou variáveis de entrada; y_i variáveis respostas ou classes utilizadas para treinamento.

Se a entrada é x podemos estimar y pelo preditor $\varphi(x_i\ ,\mathcal{L}\ )$. Agora, suponha um conjunto de preditores $\{\mathcal{L}_k\}$, cada um, com N observações independentes, originados da mesma distribuição subjacente $\mathcal{L}$, com objetivo de melhorar o aprendizado de um único $\varphi(x_i\ ,\mathcal{L}\ )$. A restrição está na autorização de trabalhar com a sequência do conjunto de preditores $\{\varphi(x_{i,},\mathcal{L}_k)\}$.

Se $\varphi(x_i\ ,\mathcal{L}\ )$ prediz uma classe $j\ \in\ \{1,\ldots,j\}$, então, um método de agregar $\varphi(x_{i,},\mathcal{L}_k)$ é pela votação da maioria. Fazer $N_j\ =\ \#\{k;\varphi(x_{i,},\mathcal{L}_k) = j\}$ e encontrar $\varphi_A(x) =\ arqmax_j\, N_j$.

Normalmente se tem apenas um conjunto de treinamento $\mathcal{L}$ sem as réplicas, que conduz ao processo de encontrar φ_A. Para isto, são efetuadas copias de amostras *bootstrap* $\{\mathcal{L}^{(B)}\}$ de $\mathcal{L}$ para $\{\varphi(x_i\ ,\mathcal{L}^{(B)}\ )\}$

Se y é uma classe, como no trabalho, pega-se $\{\varphi(x_i, \mathcal{L}^{(B)})\}$ para efetuar a votação e encontrar $\varphi_B(x)$. Este procedimento é denominado "*bootstrap aggregation*" conhecido como bagging.

Cada árvore de decisão é treinada com somente 63% das observações, por causa da escolha aleatória de n entre N observações com reposição. Essa porção dos dados é conhecida como dados "*in-bag*", enquanto os 37% de observações omitidas são as observações "*out-of-bag*". Estas últimas não são usadas para construir nem para podar qualquer árvore, mas fornecem estimativas melhores dos erros de cada nó das árvores, além de outros erros de generalização para previsores advindos de "*bagging*".

Os erros calculados das observações "*out-of-bag*" são utilizados para estimar o poder de predição e a importância dos atributos, ou variáveis de entrada. Como a habilidade de predição é mais dependente de atributos importantes e menos de atributos menos importantes, pode-se usar essa ideia para medir a importância de cada atributo. Permutando-se aleatoriamente os dados ao longo de um atributo e investigando-se o aumento do erro devido à permutação consegue-se perceber a importância desse atributo.

A técnica 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.

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

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

financeiros e a técnica logistica para antecipar falência com bons resultados, 92% e 88% respectivamente.

2. METODOLOGIA

O objetivo é construir um modelo para a previsão de insolvência das PMEs portuguesas do setor agronegócio através da metodologia ensemble denominada *tree bagging*. 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 no gráfico da 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: Descrição dos dados (ou seja, dos indicadores), limpeza dos dados, seleção das variáveis, ajuste (ou treinamento) e testes.

Na etapa Descrição dos Dados, é descrito 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.

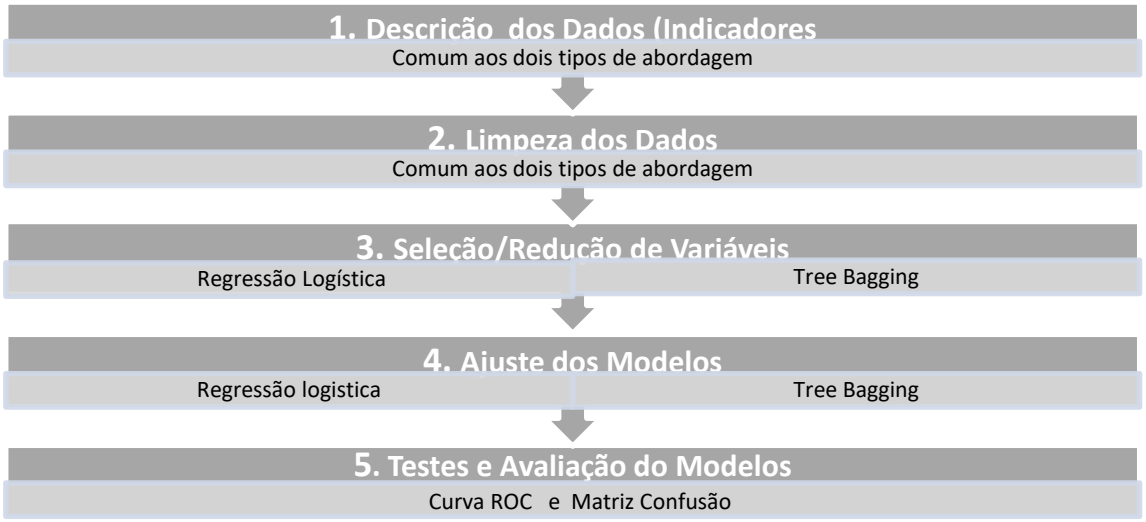

Figura 1: Metodologia experimental.

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 PME portuguesa 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 respectivos 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 PME com indicadores financeiros completos na série. As empresas foram divididas em duas classes, empresas solventes e empresas insolventes.

Critério adotado de seleção de empresa para a classe 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. O critério adotado para selecionar empresa solvente, não comtemplar capital próprio negativo no período 2007-2017.

Os critérios adotados para a escolha dos indicadores comtemplam a integridade dos dados em relação a implantação do Sistema de Normalização Contabilística em 1º 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.

Depois de selecionadas as empresas, foram selecionados 11 indicadores financeiros, conforme Tabela 1.

Tabela 1: Indicadores utilizados.

Literal	**Fórmula**
Racio de liquidez corrente	Ativo circulante / Passivo líquido
Racio de liquidez	(Ativo circulante - Estoques) / Passivo líquido
Racio de liquidez dos accionistas	Capital próprio / Passivos fixos
Racio 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
Racio 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 Própio + 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 / estoque

A limpeza dos dados é um tratamento realizado sobre os dados selecionados, de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos 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 verificado 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 gerados 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 arvores 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.

As ferramentas Matriz de Confusão e Curva ROC oferecem medidas efetivas de desempenho, ao mostrar o número o número de classificações corretas e incorretas versus as classificações preditas para cada classe em um conjunto de exemplos dicotômicos.

A Matriz de Confusão, exemplificada na tabela 2, contabiliza os dados necessários para os cálculos das métricas denominadas de acurácia, especificidade e sensibilidade.

Tabela 2: Modelo da Matriz Confusão aplicado a previsão de insolvência

Insolvente previsto	VP	FP
Solvente previsto	FN	VN
Classes	Insolvente	Solvente

VP - Verdadeiro positivo; VN - Verdadeiro negativo FP - Falso positivo; FN – Falso negativo.

O resultado FP está relacionado ao Erro Tipo I e o FN está relacionado ao Erro Tipo II

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 : $VN/(FP + VN)$. Curvas ROC ("Receiver Operating Characteristic"), conforme exemplificado na figura 1, representa a sensibilidade e a especificidade para todos os possíveis valores de pontos de corte sob a curva, será a utilizada para avaliar globalmente as metodologias utilizadas neste trabalho.

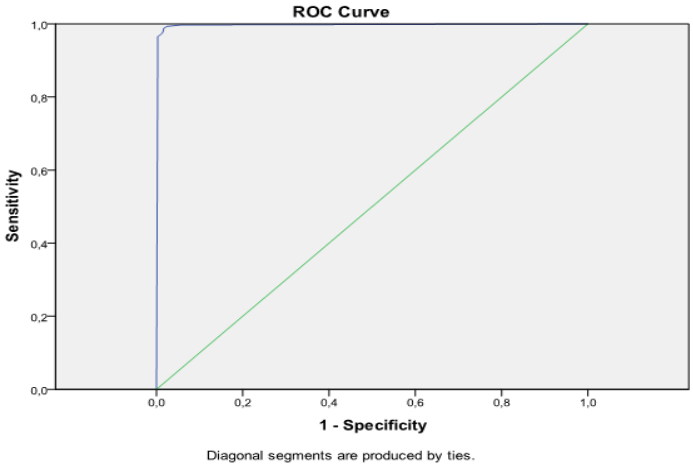

Figura 1: Exemplo de um gráfico da curva ROC extraído da plataforma SPSS

2. EXPERIMENTOS

Na base de dados inicial de 2.236 PME 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.

3.1 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
	Retorno sobre capital investido
3	Retorno sobre o total do activo
4	Margem de lucro
5	Capacidade de cobrir juros
6	Stock Turnover
7	Racio de liquidez corrente
8	Racio de liquidez
9	Racio de liquidez dos accionistas
10	Racio de solvabilidade
11	Alavancagem

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 impresso do software Matlab (Correlation Matrix)

	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	,343	0,023	-0,082	0,124	-0,155	0,114	0,471	1,000

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 respectivas PME, 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.

$$W_j = {\widehat{\beta_J}}\Big/{DP\widehat{\beta_J}}$$

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 activo e Racio 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 – Resultado adaptado do software Matlab (Estimated Coefficients)

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

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.

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.

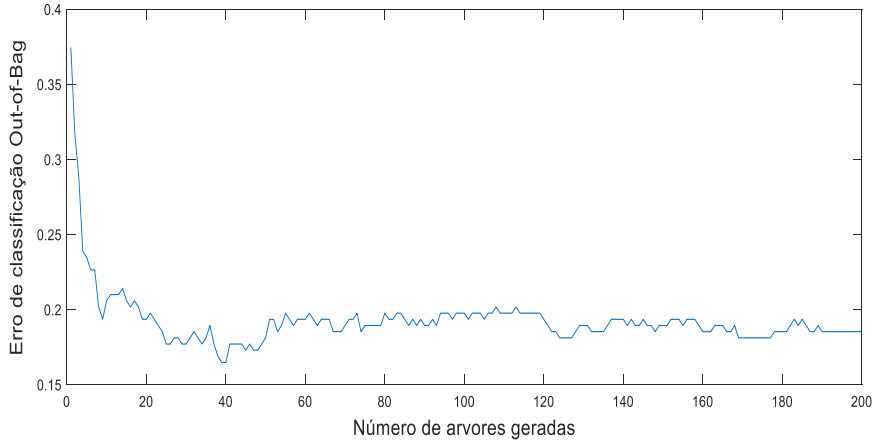

Figura 2: Variação do erro *out-of-bag* com o número de árvores geradas, Matlab.

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 .

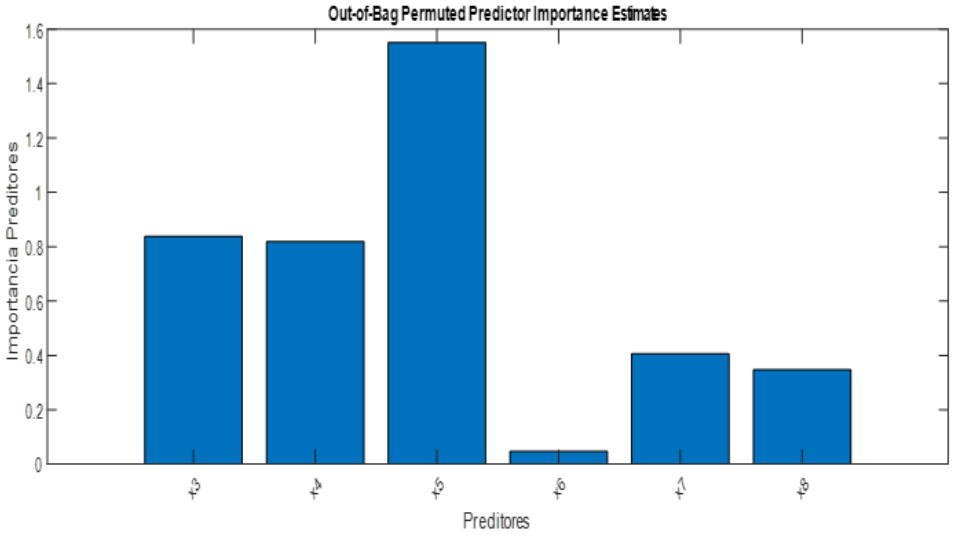

Figura 3: Importância do atributo, medida como o erro de classificação *out-of-bag*.

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.

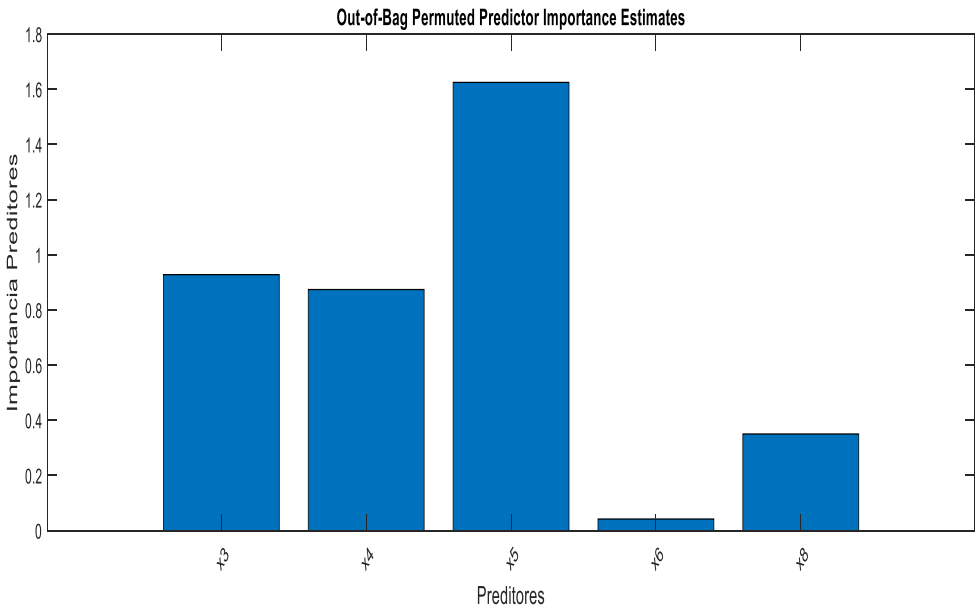

Figura 4: Importância dos atributos entre os 5 atributos selecionados.

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

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

3.2. 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 ano antes :

$$\boldsymbol{P(Y=0) = {1}/{(1 - e^{-g(x)})}}$$

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 *bagging,* 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 activo, x8 = Racio 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.

Para o ajuste foram geradas 200 árvores no conjunto com o tamanho mínimo dos nós terminais iguais a um, selecionada aleatoriamente a raiz quadrada do número total de atributos para cada divisão de decisão dos nós. O erro das observações variou com o aumento das árvores do conjunto, é esperado que esse erro fique reduzido com o número de árvores. A figura 5 apresente o gráfico da variação deste erro com o número de árvores e mostra claramente o erro diminuindo, significa que o ajuste modelo *bagging* foi adequado.

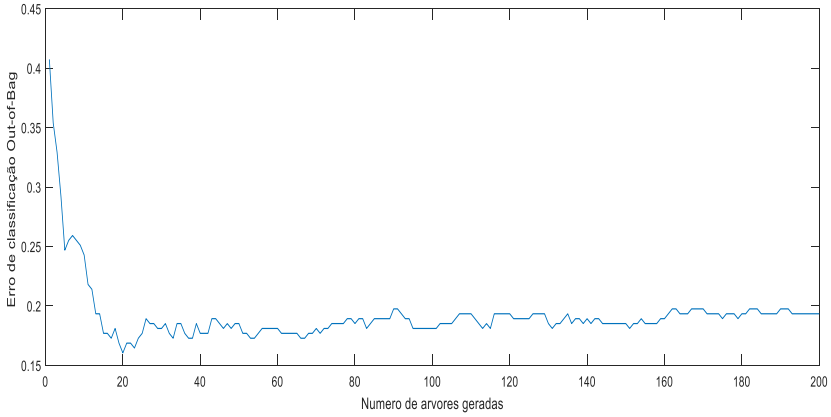

Figura 5 : Números de cópias bootsprap x erro de classificação

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 do ajuste do modelo Regressão Logística

Tabela 7: Matriz Confusão

Classe Insolvente	104	20
Classe Solvente	30	89
	0	1

Classe Predita

Tabela 8: Métricas para avaliação

Acurácia	$\frac{104+89}{243}$ $= 79{,}4\ \%$
Sensibilidade	$\frac{104}{104+30}$ $= 77{,}61\%$
Especificidade	$\frac{89}{89+20}$ $= 81{,}65\%$

As tabelas 9 e 10 apresentam os resultados do ajuste do modelo *Tree-Bagging*

Tabela 9: Matriz Confusão

Classe Insolvente	101	18
Classe Solvente	27	97
	0	1

Classe Predita

Tabela 10: Métricas de avaliação

Acurácia	$\frac{101+97}{243} =$ $81{,}48\ \%;$
Sensibilidade	$\frac{101}{101+27}$ $= 78{,}91\%$
Especificidade	$\frac{97}{97+18}$ $= 84{,}35\%$

Tabela 11: Resultados consolidados Matriz Confusão e AUC

	Acurácia	Sensibilidade	Especificidade	AUC
Regressão Logística	79,4 %	77,61%	81,65 %	0,89
Tree-Bagging	81,48 %	78,91 %	84,35 %	0,92

As métricas apresentados na tabela 11 sugerem superioridade do modelo Tree-*Bagging* em relação ao modelo tradicional de seleção Regressão Logística. O teste de Acurácia apresentou 81,48 % 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 79,4 % de probabilidade. O teste de Sensibilidade do modelo proposto apresentou 78,91% 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 84,35% de prever a solvência sendo a PME solvente, o modelo tradicional apresentou 81,65%.

O resultado 0,92 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.

4. CONCLUSÃO

Os cálculos das medidas de avaliação dos testes do modelo proposto confrontados com as medidas do modelo tradicional Regressão Logística, mais especificamente a medida de Sensibilidade, que apresenta a probabilidade de 78,91% de prever as empresas insolventes quando são insolventes sugerem a validação da metodologia *Tree-Bagging* para previsão de insolvência das PME portuguesas do setor da agroindústria com um ano de antecipação.

] Ao ampliar a análise a pesquisa acompanha o estudo de (Edmister,1972) que afirma “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”.

Como observação lateral a seleção dos indicadores mais importantes, para antecipar insolvência das PME do setor estudado sugere a necessidade do acompanhamento eficaz dos efeitos da liquidez de curto prazo, assim como no longo prazo a importância de uma relação adequada entre a capacidade de geração de resultados e os investimentos por parte das PME.

O resultado sugere também a importância no aprofundamento de estudos que utilizam a metodologia Tree-Bagging para o melhor entendimento do fenômeno insolvência para as PME portuguesas do setor agroindustrial.

REFERÊNCIAS

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

Auria, L., Moro, R. A. (2009). *Support Vector Machines (SVM) as a Technique for Solvency Analysis*. *SSRN*.
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, Sylvain; Celisse, Alain. (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. Retrieved from 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. Retrieved fromhttps://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. https://doi.org/10.1016/j.jbankfin.2016.07.015

Deng, G. (2016). *Analyzing the Risk of Mortgage Default*. University of California. Retrieved from 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 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. Retrieved from https://www.researchgate.net/profile/Jesse_Sealand/publication/326518013

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*. 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.