UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO

Relatório de Pós-Doutoramento

José Augusto Gonçalves do Canto

Orientação:

Prof. Doutor Carlos Machado dos Santos

Vila Real, 2020

LISTA DE FIGURAS

LISTA DE TABELAS

Índice

Introdução

O presente relatório atende o regulamento do programa de Pós-Doutoramento da Universidade de Trás-os-Montes e Alto Douro. O projeto apresentado à universidade em novembro de 2018 propôs elevar a compreensão do fenômeno insolvência das pequenas e médias empresas através da aplicação de algoritmos de geração de regras automáticas da metodologia de aprendizado supervisionado para antecipar um ano de antecedência a insolvência destas empresas. Dentre as opções de algoritmos foram selecionados os classificadores ensemble associados à metodologia Arvore de Decisão.

Os resultados da investigação desenvolvida no programa, abaixo descritos, sugerem acurácia consistente dos modelos propostos pela metodologia a apontam relevância dos indicadores financeiros de liquidez de curto prazo e rentabilidade do investimento como antecedentes de insolvência das pequenas e médias empresas.

1. Da problemática aos objetivos e importância da investigação

O mercado das pequenas e médias empresas (PME) possui algumas caraterísticas na qual se destaca a existência de assimetria de informações entre as PME e fornecedores de crédito, (Stiglitz e Weiss,1981) ao analisar o racionamento no mercado de crédito diagnosticam o risco moral causado pelas informações assimétricas para justificar o crédito concedido apenas a uma parcela da sua demanda. Ao concordar com este diagnóstico verifica-se a necessidade de geração e obtenção de informações mais claras e eficientes sobre a situação econômica destas empresas para atenuar o racionamento de crédito. Neste contexto o avanço tecnológico e o poder de previsão da metodologia denominada *Machine Learning (*ML*)* que serve de alternativa à tradicional comunidade estatística (metodologia que assume de maneira geral o modelo $r(x) = \beta 0 + \beta i x i$ centrado na análise dos coeficientes βi) procura melhorar a qualidade das informações sobre as PME.

Os objetivos da investigação centraram-se em elevar a compreensão do fenômeno insolvência das PME através da geração de regras automáticas que antecipe um ano antes a insolvência destas empresas. Geraram-se regras automáticas na forma de arvores através de algoritmo de aprendizado supervisionado extraídas da técnica Arvore de Decisão integrante da família ML, bastante popular pelas características: velocidade de geração, facilidade de aplicação em domínios numéricos e facilidade de compreensão para as tomadas de decisões finais. A importância da investigação assentou, ainda, na apresentação de novos instrumentos tecnológicos e acadêmicos que busquem a elevação da qualidade das informações das PME através da aplicação de uma metodologia alternativa à tradicional da comunidade estatística.

É conceptualização e a materialização destes objetivos, em concordância com o projeto apresentado, que se descreve nos pontos seguintes.

2. Atualização e amadurecimento teórico de previsão de insolvência e recursos de inteligência computacional

2.1 Insolvência: conceito

A situação de insolvência precisa ser bem caracterizada para uma melhor compreensão de sua abrangência. 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".

Independente da abordagem, a insolvência pode ter caráter apenas transitório; porém, quando o estado é resultado do capital próprio é negativo, exigem-se maiores cuidados para reverter-se a situação. Assim, este trabalho desenvolverá modelos de previsão do estado de insolvência, um ano antes de ela ocorrer.

Enquanto a origem do estado de insolvência se encontra nos prejuízos acumulados pela empresa ao longo do tempo, o conceito próximo de falência diz respeito ao processo jurídico, tomando-se o fato de que a empresa não tem como honrar os compromissos. Outro conceito originado no processo jurídico é o de concordata, o qual pressupõe que, através da concessão de maior prazo para pagamento de seus compromissos contratuais ou redução de montante destas, ou ambos, possa recuperar sua solvência.

Isso significa que os modelos de previsão podem ser elaborados tendo em vista empresas com a quebra decretada em tribunal (falência ou concordata) ou, de forma diferente, em diversos conceitos de insolvência ou dificuldade financeira, situação em que a empresa apresenta valor de capital próprio negativo.

2.2 Evolução dos modelos de previsão de insolvência

Nos anos 30, foram desenvolvidos os primeiros estudos sobre previsão de insolvência, mas foi Beaver (1966) que apresentou o primeiro trabalho com técnicas estatísticas ao empregar dados contábeis para prever falência, através de discriminação univariada.

Altman (1968) deu impulso aos estudos dos modelos de previsão, apesar do resultado de sua função discriminante, denominado de Z Score ser um número pouco intuitivo. Desse momento em diante, a quantidade de pesquisas para tratar do problema da previsão de insolvência, falência e dificuldades financeiras cresceu em todo o planeta e, de acordo com Kumar e Ravi (2007) de especial significado deve-se destacar o que elas têm em comum: utilização de conjuntos de indicadores financeiros no país de origem da pesquisa como fonte de dados; preocupação em definir a linha do tempo do conjunto de dados; estudo comparativo das técnicas quanto ao desempenho em termos de precisão da previsão.

A função logística é um tipo especial de função de distribuição acumulada, sendo identificada, mas precisamente, como função de distribuição acumulada logística.

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

A análise logística de Ohlson (1980) utilizou oito indicadores financeiros e foi capaz de identificar, com um ano de antecedência, a falência de empresas com 89% de precisão. Platt

e Platt (1991), ao elaborarem seus modelos, recomendam a utilização de índices financeiros padronizados pelo setor, no local de indicadores absolutos das empresas.

A utilização de indicadores financeiros extraídos da contabilidade versus mercado de capitais foi causa de publicação de diversos artigos. Shumway (2001) aborda os benefícios da inclusão de indicadores financeiros captados no mercado, nos modelos para a previsão de falência, e concluí que os índices financeiros de mercado apresentam baixa correlação com os índices financeiros retirados da contabilidade e melhoram a precisão dos modelos que utilizam exclusivamente os dados contábeis.

Paralelamente ao crescimento dos trabalhos que combinavam os indicadores financeiros à estatística clássica, o avanço da tecnologia computacional propiciou o surgimento de modelos à base de algoritmos de aprendizado de máquina, desenvolvidos com recursos da inteligência computacional. Apesar de ambas estudarem a análise de dados, diferentemente da estatística, as técnicas de aprendizado em máquina não focam em modelos teóricos bem definidos à procura de ajustamento de parâmetros a esses modelos teóricos, e sim em algoritmos que utilizam modelos mais flexíveis e heurísticas para a realização de busca. Uma análise estatística dos dados pode determinar as tendências centrais, variabilidades e as correlações entre atributos que descrevem os fatos, mas não é capaz de caracterizar a relação destes atributos em um nível abstrato e conceitual, nem descrever a relação causal caso exista.

Embora a utilização dos algoritmos de aprendizado de máquina nos trabalhos de finanças seja relativamente nova a evolução tecnológica 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. Trabalhos têm vindo a ser 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), Bagherpour (2017) , Addo et al., (2018) e Tokpavi (2018) comparam bons resultados com o modelo estatístico

tradicional Regressão Logística. Dentre as opções (Quinlan, 1986) já destacava a maior facilidade de compreensão do algorítmo Árvores de Decisão por ser fortemente intuitivo.

2.2.1 Algorítmo Arvore de Decisão

Quinlan (1986), que popularizou o sistema indutivo de regras, denominado árvores de decisão, apresenta a técnica como algoritmos indutivos utilizados por programas de computadores que pertencem à família TDIDT ("*Top Dow Induction of Decision Trees"*). Explica que a ideia básica dos algoritmos surge da estratégia "dividir-para-conquistar"; são divisões recursivas do espaço dos dados de treinamento, cuja representação simbólica das partições pode ser uma árvore invertida. No estudo dos algoritmos, o conjunto dos dados de treinamento é constituído por objetos (exemplos) descritos por atributos que são variáveis observáveis e independentes, associados a uma variável dependente, denominada de classe.

O algoritmo da árvore de decisão busca meios de dividir um conjunto de dados em vários subconjuntos disjuntos, conhecidos simbolicamente como nós. Cada nó representa um conjunto composto por exemplos (registros) os quais atendem a um teste sobre um valor específico assumido por um dos atributos dos exemplos. O algoritmo é baseado na técnica de particionamento recursivo.

Quinlan (1986) explica, através de um exercício, a técnica de indução de uma árvore. A base da indução em um problema de classificação é o universo de objetos, descrito a partir de um conjunto de atributos. Cada atributo mede alguma característica importante de um objeto. Por exemplo, se os objetos são manhãs de sábado e a tarefa de classificação envolve o tempo, os atributos poderiam ser:

- Tempo, com valores {ensolarado, nublado, chuva}
- Temperatura, com valores {fria, moderada, quente}
- Umidade, com valores {alta, normal}
- Ventania, {verdadeiro, falso}

Tomados em conjunto, os atributos que caracterizariam os objetos no universo “manhã de sábado” poderiam ser assim descritas:

Tempo: nublado;

- Temperatura: frio;
- Umidade: normal;
- Ventania: falso.

Se cada objeto do universo vier a pertencer a duas classes denotadas de *P* e *N*, referidas como positivo e negativo, formará um conjunto de treinamento. A tarefa de indução é o desenvolvimento de uma regra de indução que possa determinar exclusivamente a classe *P ou N* de qualquer objeto, em função dos valores de seus atributos. A questão é se o conjunto de treino, através de seus atributos, fornece informações suficientes para isso. Se o conjunto de treino contém dois objetos que têm valores idênticos para cada atributo e ainda pertençam a diferentes classes, é claramente impossível diferenciar entre esses objetos apenas com os atributos dados. Em tal caso, os atributos são denominados inadequados para o conjunto de treino e, consequentemente para a tarefa de treino.

A tabela 1 mostra um pequeno conjunto de treinamento que utiliza os atributos da “manhã de sábado”. Cada objeto e seus atributos são mostrados em conjunto com a classe (aqui, a classe P manhãs é adequada para alguma atividade não especificada, enquanto a classe N manhãs não é adequada para alguma atividade não especificada).

Tabela 1: Pequeno conjunto de treinamento.

NÚMERO	**ATRIBUTOS**				**CLASSE**
	Tempo	**Temperatura**	**Umidade**	**Ventania**	
01	ensolarado	quente	alta	falso	N
02	ensolarado	quente	alta	verdadeiro	N
03	nublado	quente	alta	falso	P
04	chuvoso	moderada	alta	falso	P
05	chuvoso	frio	normal	falso	P
06	chuvoso	frio	normal	verdadeiro	N
07	nublado	frio	normal	verdadeiro	P

08	ensolarado	moderada	alta	falso	N
09	ensolarado	frio	normal	falso	P
10	chuvoso	moderada	normal	falso	P
11	ensolarado	moderada	normal	verdadeiro	P
12	nublado	moderada	alta	verdadeiro	P
13	nublado	quente	normal	falso	P
14	chuvoso	moderada	alta	verdadeiro	N

A árvore de decisão que classifica os objetos do pequeno conjunto de treinamento, contidos na tabela 3, é mostrada na Figura 1. As folhas da árvore representam as classes, outros nós representam os atributos testados com um ramo para cada resultado possível.

A fase de treinamento na classificação geralmente segue o roteiro: a) uma das características dos atributos é escolhida segundo algum critério de seleção; tal critério é usado para particionar o conjunto completo de exemplos; b) a cada possível resultado de avaliação do critério, agrupam-se objetos que atendam a tal condição, formando subconjuntos disjuntos de exemplos; c) para cada subconjunto, o processo é repetido até que existam apenas exemplos do mesmo tipo em cada subconjunto resultante.

Os critérios de seleção definem para cada nó da árvore o atributo a ser utilizado na construção da árvore. Tan, Steinbach e Kumar (2005) relatam que existem diferentes tipos de critérios de seleção, utilizados em diversos algoritmos de indução de árvores de classificação. De maneira geral, os algoritmos de indução buscam dividir os dados de um nó interno, baseados em um único atributo, com a utilização de medidas para tentar-se encontrar o melhor atributo para efetuar essa divisão. A maior parte dos algoritmos utiliza o grau de impureza como medida: quanto menor, maior é a distribuição de classes. O grau de impureza é máximo no nó, se houver o mesmo número de exemplos para cada classe possível; é nula se todos os exemplos nele pertencerem à mesma classe.

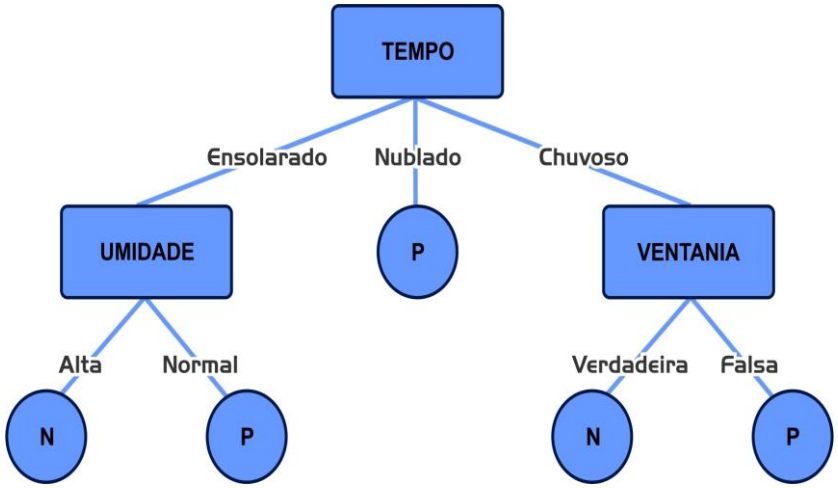

Figura 1: Exemplo de árvore de decisão

No processo de construção, a árvore de classificação utiliza uma função conhecida como função-impureza que irá procurar exaustivamente, por meio de um processo recursivo, minimizar a margem de erro. A margem de erro, para um dado nó, é mínima quando todos os dados pertencem à mesma classe, e máxima quando os dados são linearmente distribuídos através das classes.

As funções-impurezas – Função Entropia e Gini Index – são citadas em Tan, Steinbach e Kumar (2005) como as mais utilizadas para uma árvore de classificação.

$$Entropia(N) = \sum_{j=1}^{m} -p_j \, log_2 p_j \qquad (4)$$

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

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, através da seleção de atributos que melhor se enquadram no nó em análise, para posterior avaliação da melhor partição. Uma das medidas s conhecidas de seleção de atributos é o Ganho de Informação, utilizado no algoritmo ID3, para determinar a qualidade do teste realizado,

compara o grau de entropia antes da divisão com o grau de entropia depois da divisão. O atributo que apresentar a maior diferença é escolhido, formalmente:

$$\Delta Ganho(N,t) = Entropia(N) - Entropia(N_l) - Entropia(N_r) \qquad (5)$$

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 atributos que possuem muitos valores, Quinlan (1986) propõem uma alternativa de seleção de atributos baseada na Razão do Ganho de Informação. O critério Ganho de informação apresenta uma desvantagem de tender a ser muito grande quando os atributos possuem muitos valores, pois geram árvores muito largas.

$$RGanho(S,A) = \frac{Ganho(S,A)}{Info(S,A)}$$

$$info(S,A) = -\sum_{i}^{n} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

Onde S_i é o subconjunto de exemplos após o particionamento de S pelos *n* valores do atributo *A.* Os programas de indução ID3 adotam um método que depende de duas premissas. Seja: *C* contêm *v* objetos das classes *P* e *N*. As hipóteses são:

(1) Qualquer árvore de decisão correta para *C* irá classificar os objetos na mesma proporção de sua representação em *C.* Um objeto arbitrário será determinado a fazer parte de classe P com probabilidade p / (p + n) e probabilidade n / (p + n) para a classe N.

(2) Quando uma árvore de decisão é usada para classificar um objeto, ele retorna uma classe. A árvore de decisão pode, assim, ser considerada como fonte de uma mensagem “P”

ou “N”, com a informação necessária para gerar mensagem esperada, dada pela equação, conforme a função de impureza 4:

$$I\,(p,n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}$$

Se um atributo com valores [A_I, A_2. . .. A_v] é usado para a raiz da árvore de decisão, e partição C em [C_1, C_a. . . . C_v] onde C_i contém esses objetos em C e A_i em A. Ci vai conter objetos pi de classe P e ni da classe N. As informações necessárias para a sub árvore esperada de Ci é I (pi, ni). A informação necessária para a árvore esperada com o atributo A como raiz é então obtido como a média ponderada

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p,n) \qquad (6)$$

Onde o peso para o ramo é a proporção dos objetos em C que pertencem a Ci. A informação obtida através de ramificação em A é, por conseguinte, $Ganho(A) = I(p,n) - E(A)$.

Uma boa regra parece ser a de escolher esse atributo para o ramo em que ganha a maioria das informações. O ID3 examina todos os atributos candidatos e escolhe um para maximizar o ganho (A), forma a árvore, e usa o mesmo processo recursivamente para formar árvores de decisão, para o residual dos subconjuntos C_1, C_E. . . . C_v.

Para ilustrar a ideia, seja *C* o conjunto de objetos na Tabela 3. Dos quatorze objetos, nove são de classe P e cinco são de classe N, então a informação necessária para a classificação é

$$I(p,n) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0{,}94\ bits$$

Se for considerado o atributo *tempo* com valores {*ensolarado; nublado; chuva*}, cinco dos quatorze objetos em C tem o primeiro valor (ensolarado), dois deles de classe P e três da classe N, então:

$$p_1 = 2\,, n_1 = 3 \quad I(p_1, n_1) = 0{,}971$$

E similarmente:

$$p_2 = 4\,, n_2 = 0 \quad I(p_2, n_2) = 0$$

$$p_3 = 3, n_3 = 2 \quad I(p_3, n_3) = 0{,}971$$

A exigência de informações que se espera depois de testar este atributo é, portanto,

$$E(tempo) = \frac{5}{14} I(p_1, n_1) + \frac{4}{14} I(p_2, n_2) + \frac{5}{14} I(p_3, n_3) = 0{,}694\ bits$$

Como l (p, n) é constante para todos os atributos, maximizar o ganho é equivalente a minimizar E(A). O ganho desse atributo, conforme equação 5, é então: Ganho(*tempo*) = 0.940 – E(*tempo*) = 0.246 bits.

Similarmente, há: Ganho (temperatura) = 0,029 bits; Ganho (umidade) = 0,151 bits e Ganho (vento) = 0,048 bits, de modo que o método de árvore de formação, utilizado em ID3, escolheria o tempo como o atributo para a raiz da árvore de decisão. Os objetos, em seguida, serão divididos em subgrupos de acordo com os valores de seus atributos, e uma árvore de decisão para cada um dos subconjuntos será induzida em uma forma similar. De fato, a Figura 2 mostra a árvore de decisão real gerada por ID3, a partir desse conjunto de treinamento.

Apesar de muito popular a Árvore de Decisão a presente o 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, as Árvores de Decisão são modelos instáveis, pequenas variações nos dados de entrada podem resultar árvores completamente distintas. Para evitar a alta variância foi desenvolvido técnicas ensemble para gerar várias árvores distintas com agregação de resultados. Para este processo de agregação utiliza-se os processos ensemble *bagging* ou *boosteed.*

O termo "*bagging"* vem de "*bootstrap aggregation"*. *Bagging* de preditores é um método para gerar versões múltiplas de um preditor e utilizar as versões com vistas a ter um preditor agregado. A agregação combina os resultados de todas as versões, normalmente

fazendo uma média ponderada, para a predição de uma saída numérica e, em geral, faz um voto de pluralidade quando o problema é prever uma classe. No trabalho foi utilizado a técnica Tree-Bagging, se constrói múltiplas arvores, réplicas "bootstrap", para compor o conjunto de treinamento (ou de ajuste), cada réplica de arvore funciona como um classificador treinado, o conjunto de réplicas gera um comitê de arvores, que através do voto prevê um novo dado.

No processo *boosteed* 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. No trabalho foi utilizado a técnica denominada de Adaboost, se refere a um método específico de treinamento de um classificador aprimorado, que utiliza uma combinação dos processos ensemble b*agging* e *boosted* para melhorar o desempenho do algoritmo de aprendizado supervisionado Arvore de Decisão, no método as réplicas de arvores funcionam como classificadores fracos que convergem a um classificador forte.

2.2.2.Técnica ensemble Tree Bagging

A técnica T*ree 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.

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.

Na metodologia *bagging* cada réplica de arvore funciona como um classificador treinado, o conjunto de réplicas gera um comitê de arvores, 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.

2.2.3. Técnica ensemble *Adaboost*

A segunda metodologia proposta se refere a um método específico de treinamento de um classificador aprimorado, que utiliza o algorítmo *AdaBoost,* "*Adaptive Boosting*" uma combinação das ideias de b*agging* e b*oosting* para melhorar o desempenho do algoritmo de aprendizado supervisionado Arvore de Decisão. No método réplicas de arvore funcionam como classificadores fracos que convergem a um classificador forte. Conforme (Sahapire, 1990) no algorítmo *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:

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

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 $\in_t$.

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

$F_{t-1}(x_i)$ 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 arvores com pesos altos.

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

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 $\in_t$.

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

$F_{t-1}(x_i)$ 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 arvores 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:

$$\in_t = \Pr[\mathrm{ht(xi)}$$

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

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

3. Construção de indicadores de entrada, seleção das variáveis, ajustes e validação dos modelos

3.1. Metodologia

No planejamento metodológico apresentado no projeto seria utilizada apenas a técnica ensemble *bagging* para efetuar o ajuste do modelo de previsão, porém ao longo da revisão bibliográfica verificou-se que a técnica ensemble *AdaBoost* que introduz voto qualificado no comitê de máquinas é muito interessante. Por esta razão aplicaram-se as duas técnicas ensemble para construir dois para a previsão de insolvência das PME portuguesas do setor agronegócio. A validação dos modelos propostos segue a metodologia de trabalhos correlacionados ao utilizar um modelo estatístico tradicional como parâmetro de desempenho, no projeto foi previsto utilizar a técnica de seleção linear Análise Descriminante Linear, porém foi verificado na fase de revisão bibliográfica a grande utilização da técnica de seleção Análise de Regressão Logística para validação, o que acarretou mudança.

A regressão logística é uma técnica em que a variável dependente é de natureza dicotômica ou binária, atribuindo-se o valor 1 ao sucesso do interesse e 0 ao insucesso, bastante utilizada na comunidade estatística para classificar se a empresa encontra-se no grupo de empresas solvente ou insolvente, utiliza a função de distribuição logística para modelar dados, descrita por (Zavgren,1985) como um tipo especial de função, sendo identificada mais precisamente como função de distribuição acumulada logística.

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

Os experimentos realizados são divididos em dois grupos: ajustes e testes com modelagem logística e ajustes e testes com modelos ensemble. Por serem distintas as metodologias elas serão realizadas separadamente, tendo em comum somente a primeira fase. A metodologia está dividida em 3 fases: seleção dos dados (ou seja, indicadores financeiros), seleção das variáveis (exemplos para o treinamento)/ ajuste (ou treinamento) e testes.

3.2. Descrição dos dados

3.2.1. Indicadores financeiros

Assume-se o pressuposto de acumulação de informações nas demonstrações contábeis e concorda com Beaver (2006), ao argumentar que o mesmo irá ocorrer com os indicadores decorrentes de processos aritméticos entre tais demonstrações, portanto, pode-se esperar que os indicadores tenham um poder explicativo e informacional sobre as empresas, por serem uma sucessão de todos os eventos econômicos ocorridos nelas que justifica o seu uso como preditores ou estimadores da probabilidade de falência das empresas, ou seja:

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

No processo de análise tradicional, os indicadores são comparadores de duas formas: em cortes transversais ou de séries temporais. A análise em corte transversal envolve a comparação de indicadores de diferentes empresas na mesma data "c*ross- section*" , geralmente para avaliar o desempenho de uma empresa em relação a outras. A análise de séries temporais avalia o desempenho com o passar do tempo, ao comparar o presente com o desempenho passado. Análise "*cross-section"* foi escolhida por ser adequar melhor ao treinamento.

Incialmente foi planejado utilizar os indicadores financeiros de PME portuguesas e espanholas, mas o acesso das espanholas não teve êxito. Assim a base de dados utilizada contém indicadores financeiros europeus das PME portuguesas contidas 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 PME organizadas em "*cross-section"* observou o intervalo temporal 2008-2017 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 2008-2017.

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 2.

Tabela 2: 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

3.3. Seleção/reução das variáveis/ajustes/testes

A seleção das variáveis será feita separadamente para a modelagem logística e para as modelagens ensemble. Para selecionar as variáveis da modelagem Regressão Logística é aplicado o teste paramétrico Wald para analisar os valores dos coeficientes gerados e verificar a hipótese nula ao nível de 5%

As variáveis do modelo proposto são extraídas automaticamente por uma técnica baseada em modelagem de árvores de decisão, a técnica "*tree bagging"*– de Breiman (1996b) e de Sutton (2005), que associa o processo de "*bagging"* às árvores de decisão – foi utilizada para a seleção de variáveis dos modelos ensemble.

A seleção do conjunto de treinamento pode trazer mudanças significativas no preditor construído, ou seja, os modelos gerados podem diferir muito, e o processo de *bagging* pode melhorar a acurácia. O elemento vital é a instabilidade do previsor. Se o tipo de preditor estiver sujeito à instabilidade, os resultados poderão ter a acurácia significativamente melhorada através do "*bagging"*.

Portanto, o processo de "*tree bagging"* identifica o treinamento de árvores de decisão em réplicas de "*bootstrap"* dos dados de entrada. Essas réplicas são obtidas pela seleção

aleatória de n entre N observações com reposição, onde N é o tamanho do conjunto de dados. Para cada réplica é gerada uma árvore de decisão, a qual funciona como um classificador treinado e pode ser usada isoladamente para classificar novos dados.

As predições de duas árvores treinadas, a partir de duas diferentes réplicas de "*bootstrap"*, podem ser diferentes. O conjunto de árvores agrega as predições de todas as árvores de decisão que são geradas de todas as réplicas. Utilizando-se o critério de maioria, se para um novo dado é prevista uma determinada classe na maioria das árvores, é razoável supor que essa predição é bem mais robusta que a predição de uma única árvore. Além disso, se uma classe diferente é prevista por um conjunto menor de árvores, esta informação também é útil. A proporção do número de árvores que predizem diferentes classes serve como base para gerar scores de classificação. Em um problema de regressão, a resposta prevista de um conjunto de árvores treinadas normalmente é média das predições das árvores individuais.

Cada árvore de decisão é treinada com somente 63%, em média, 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"*. As observações podem ser usadas para estimar o poder de predição e a importância dos atributos, ou variáveis de entrada. A predição para cada observação é estimada pela média das predições de todas as árvores no conjunto para as quais a observação é "*out-of-bag"*. O erro médio "*out-of-bag"* é um estimador não tendencioso do erro real do conjunto.

A habilidade de predição depende mais 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, ou seja, para cada atributo permutam-se os valores dele em todas as observações do conjunto de dados e mede-se o quão pior o erro médio quadrático (MSE) ficou depois da permutação.

Armazenando-se o aumento do MSE, ponderado sobre todas as árvores do conjunto e divididos pelo desvio padrão calculado considerando-se todas as árvores, para cada atributo pode-se quantificar sua importância, em separado. Quanto maior o valor, mais importante é a variável.

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.

As técnicas de validação cruzada) consistem em dividir o conjunto de dados em subconjuntos mutuamente exclusivos e em utilizar alguns dos subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) na validação do modelo.

O método de validação cruzada denominado *k-fold* consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste, os k-1 restantes são utilizados para estimação dos parâmetros e calculada a acurácia do modelo. O processo é realizado k vezes, alternando-se, de forma circular, o subconjunto de teste. Ao final das k iterações, calcula-se a acurácia sobre os erros encontrados, obtendo-se, assim, uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados.

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.

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 3, contabiliza os dados necessários para os cálculos das métricas denominadas de acurácia, especificidade e sensibilidade.

Tabela 3: Modelo da Matriz Confusão

Insolvente previsto	VP	FP	VP - Verdadeiro positivo; VN - Verdadeiro negativo FP - Falso positivo; FN – Falso negativo.
Solvente previsto	FN	VN	O resultado FP está relacionado ao Erro Tipo I e o FN está relacionado ao Erro Tipo II
Classes	Insolvente	Solvente	

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. 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.

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 4: 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	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 5 – 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 conter o atributo capital próprio nas suas construções.

4.2. Seleção das variáveis para a modelagem logística

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 6 – 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
	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

4.3. Seleção das variáveis para as modelagens ensemble

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.

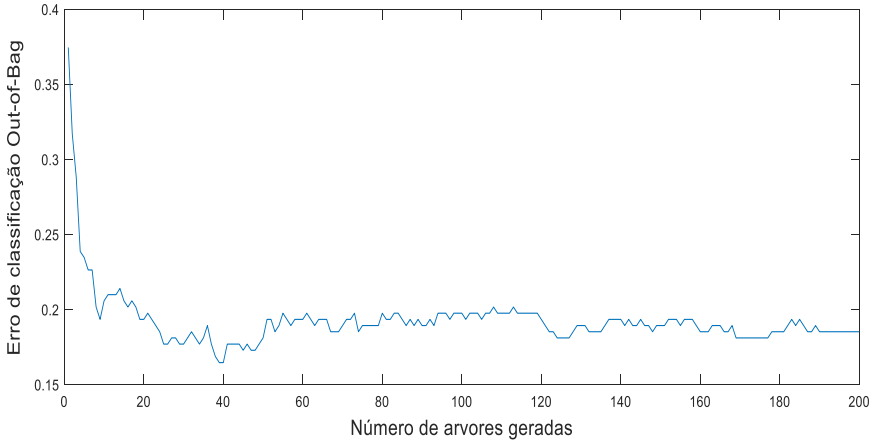

Figura 2: Variação do erro *out-of-bag* com o número de árvores geradas, Matlab.

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.

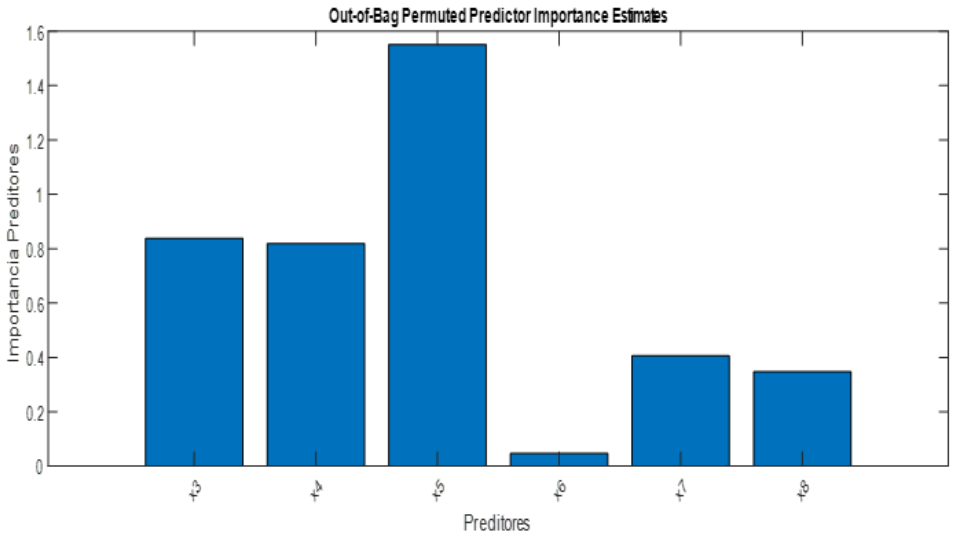

Figura 3: Importância do atributo, medida como o erro de classificação *out-of-bag*.

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.

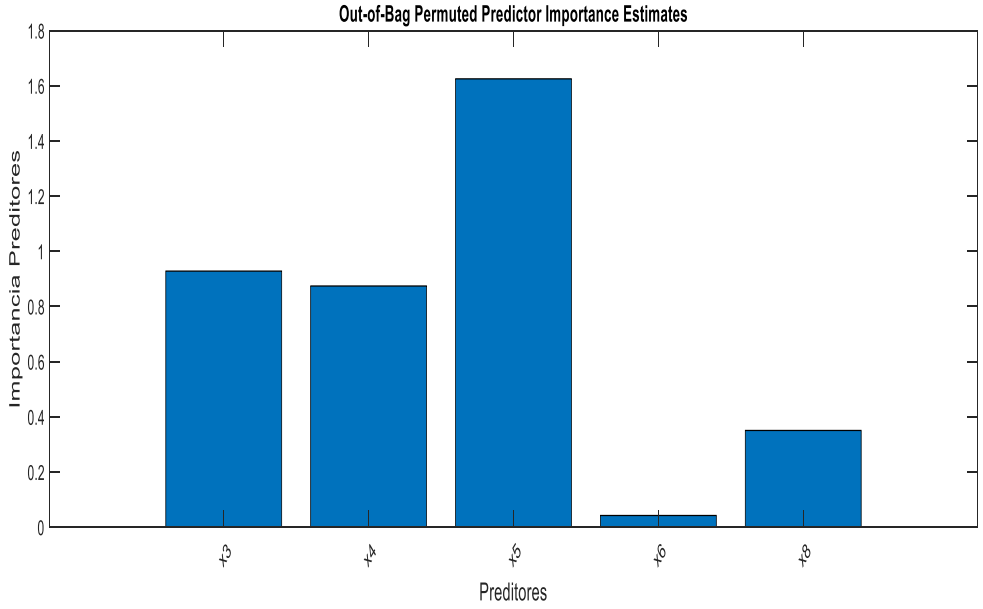

Figura 4: Importância dos atributos entre os 5 atributos selecionados.

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 7: 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.

4.4. Ajuste dos Modelos

4.4.1. Ajuste do modelo logístico

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 :

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

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

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

O ajuste do modelo logístico para as PME portuguesas do setor agroindustrial demonstra a importância do retorno sobre o total dos valores investidos em ativos e a liquidez de curto prazo como preditores de insolvência. 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 8 e 9 apresentam os resultados do ajuste do modelo Regressão Logística

Tabela 8: Matriz Confusão Logística

Insolvente	104	20
Solvente	30	89
	0 Classe	1 Predita

Tabela 9: Métricas para avaliação Logística

Acurácia	-	$\frac{104+89}{243}$ 79,4 %
Sensibilidade	-	$\frac{104}{104+30} = 7{,}61\%$
Especificidade	-	$\frac{89}{89+20} = 81{,}65\%$

4.4.2. Ajuste do modelo Tree-Bagging

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 ativo, 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 Tree-B*agging* foi adequado.

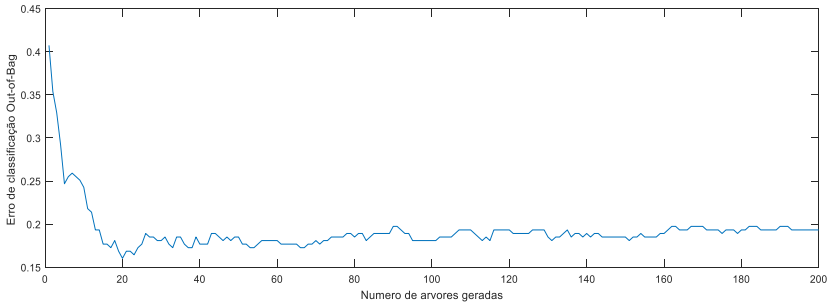

Figura 5 : Números de cópias bootsprap x erro de classificação

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

Tabela 10: Matriz Confusão Tree-Bagging

	101	18
Insolvente		
	27	97
Solvente		
	0	1
	Classe	Predita

Tabela 11: Métricas para avalição Tree-Bagging

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

4.4.3. Ajuste do modelo Adaboost

Na metodologia Adaboost a base do sistema proposto também 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, as tabelas 9 e 10 apresentam os resultados do ajuste do modelo Adaboost. O ajuste do modelo Adaboost seguiu os passos:

1) Atribuir a todos exemplos observados x_i um peso inicial $w_i = 1/n$;
2) Treinar uma Arvore de Decisão;
3) Para cada exemplo observado:
   3.1) Se predição for incorreta w_i é incrementado pelo valor parametrizado 1;
   3.2) Se predição for correta w_i é decrementado pelo valor parametrizado -1;
4) Treinar uma nova Arvore de Decisão dando prioridade aos exemplos de maior w_i
5) Repetir os passos 3 e 4 até alcançar o número parametrizado de 200 arvores, as preditoras com menos de 50% de acerto são abandonadas.

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 12 e 13 apresentam os resultados do ajuste do modelo Adaboost.

Tabela 12: Matriz Confusão Adaboost

Insolvente	90	29
Solvente	19	105
	0	1
	Classe	Predita

Tabela 13: Métricas para avalição Adaboost

Acurácia -	$\frac{101 + 97}{243}$ 80,25 %
Sensibilidade -	$\frac{90}{90 + 19} = 82{,}57$
Especificidade -	$\frac{105}{105 + 29} = 78{,}36$

4.5. Avaliação dos resultados

Para avaliar os resultados da pesquisa utilizaram-se as métricas de acurácia, sensibilidade, especificidade e usamos a curva ROC (Área sob a curva) e a AUC (Características Operacionais do Receptor), importantes para verificar o desempenho de qualquer modelo de classificação.

Tabela 14: 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
Adaboost	80,25	82,57	78,36	0,90

As métricas apresentados na tabela 14 sugerem superioridade dos modelos *Tree-Bagging e Adaboost* em relação ao modelo tradicional de seleção Regressão Logística: Os modelos propostos apresentaram no teste de Acurácia as probabilidades de 81,48 % e 80,25 % respectivamente de acertarem a previsão do estado de insolvência das PME portuguesas do setor agroindustrial com um ano de antecedência, enquanto o modelo tradicional 79,4 % de probabilidade.

Nos testes de Sensibilidade apresentaram 8,91% e 82,57% respectivamente a probabilidade de preverem a insolvência com um ano de antecedência sendo a PME na situação de ser realmente insolvente, o modelo tradicional 77,61 %. No teste de Especificidade os modelos propostos apresentaram as probabilidades de 84,35% e 78,36 % respectivamente de prever a solvência sendo a PME solvente, o modelo tradicional apresentou 81,65%.

As metodologias propostas apresentaram os resultados 0,92 e 0,90 no teste de ajuste da medida AUC da curva ROC e 0,89 na metodologia tradicional, o indica qualidade superior

do ajuste das metodologias propostas para prever insolvência das PME quando o ponto de corte das medidas de sensibilidade e especificidade são expandidos.

Conclusões

Com objetivo de aprofundar a compreensão do fenômeno insolvência das PME portuguesas do setor agroindustrial através da utilização de modelos de aprendizagem de máquina, selecionaram-se as metodologias ensemble *Tree-Bagging* e *Adaboost.* As conclusões atingidas foram as seguintes:

1) Os cálculos das medidas de avaliação dos testes dos modelos Ensemble confrontados com as medidas do modelo tradicional Regressão Logística, especificamente a medida de Sensibilidade, sugerem a superioridade das metodologias propostas para predizer insolvência das PME portuguesas.

2) A relevância dos indicadores financeiros tradicionais de liquidez de curto prazo e retorno sobre investimento como variáveis preditoras para as PME portuguesas do setor de agroindustrial, os dois indicadores foram selecionados na metodologia tradicional e na metodologia ensemble para antecipar insolvência.

A análise dos resultados da investigação recomenda a utilização da metodologia ensemble para aprofundar o estudo do fenômeno insolvência das PME portuguesas e o acompanhamento da liquidez de curto e o retorno sobre o investimento para a saúde financeira destas empresas.

Bibliografia

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.

Stiglitz, J., Weiss, A. (1981) Credit rationing in markets with imperfect information. American Economic Review, v. 71, p. 393-410.

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.

Outros

I. Participação em eventos

Canto, J.A.; Silva, A.; Leite, G.; Machado-Santos, C. (2019). “Previsão de insolvência das PME portuguesas do setor agroindustrial: Metodologia Tree-Bagging”. Comunicação apresentada no XVII Congresso Internacional de Contabilidade e Auditoria (CICA). Porto.

Canto, J.A.; Silva, A.; Leite, G.; Machado-Santos, C. (2020). “Uma aplicação de Machine Learning para previsão de insolvência: Metodologia Adaboost”. Comunicação apresentada nas XXXX Jornadas Luso-Espanholas de Gestão Cientifica. Bragança.

II. Publicações

Canto, J.A.; Silva, A.; Leite, G.; Machado-Santos, C.: "Insolvency prediction for Portuguese agro-industrial SME: Tree Bagging Methodology". Aceite para publicação na revista Agricultural Economics Review (AER2019). Junho de 2020.

Canto, J.A.; Silva, A.; Leite, G.; Machado-Santos, C. (2019). "Uma aplicação de Machine Learning para previsão de insolvência: Metodologia Adaboost". Selecionado para publicação nas XXXX Jornadas Luso-Espanholas de Gestão.

Anexos

Anexo 1: "Insolvency prediction for Portuguese agro-industrial SME: Tree Bagging Methodology".

Anexo 2: Aceite para publicação na revista Agricultural Economics Review (AER2019) – junho de 2020

Anexo 3: Selecionado para publicação nas XXXX Jornadas Luso-Espanholas de Gestão (2019) "Uma aplicação de Machine Learning para previsão de insolvência: Metodologia Adaboost".

Anexo 4: Comunicação apresentada no XVII Congresso Internacional de Contabilidade e Auditoria (CICA) – Porto. "Previsão de insolvência das PME portuguesas do setor agroindustrial: Metodologia Tree-Bagging".