UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMICA, ADMINISTRAÇÃO, ATUÁRIA,
CONTABILIDADE E SECRETARIADO
CURSO DE CIÊNCIAS ATUARIAIS

RÔMULO ALVES SOARES

MODELOS DE CLASSIFICAÇÃO APLICADOS À PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO

FORTALEZA
2013

RÔMULO ALVES SOARES

MODELOS DE CLASSIFICAÇÃO APLICADOS À PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO

Monografia apresentada ao Curso de Ciências Atuariais da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Orientador(a): Profª. Drª. Sílvia Maria Dias Pedro Rebouças.

FORTALEZA

2013

RÔMULO ALVES SOARES

MODELOS DE CLASSIFICAÇÃO APLICADOS À PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO

Monografia apresentada ao Curso de Ciências Atuariais da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Aprovada em: ____ / ____ / _____.

BANCA EXAMINADORA

_____

Profª. Drª. Sílvia Maria Dias Pedro Rebouças (Orientadora)
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Antônio Carlos Dias Coelho
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Vicente Lima Crisóstomo
Universidade Federal do Ceará (UFC)

À minha família.

AGRADECIMENTOS

À minha professora e orientadora, Professora Sílvia Rebouças, primeiramente pela confiança, paciência e dedicação sempre demonstradas, mas acima de tudo por ser um exemplo de profissional e grande educadora.

Ao João Felipe, pelo auxílio prestado à execução deste trabalho, sempre muito prestativo ao disponibilizar os dados necessários.

Aos meus colegas de curso, com quem partilhei muitos momentos ao longo do curso de Ciências Atuariais. Em especial à minha amiga Michele, que me acompanhou de perto nesses momentos derradeiros, dividindo momentos de angústia e incerteza. Agora poderemos celebrar nossa vitória.

Àqueles que sempre se revelaram bons amigos ao longo dos anos. Em especial à minha tão querida amiga Gabriele, que com sua humanidade, inteligência e teimosia, foi uma pessoa que me fez mudar muito nos últimos anos, e por isso eu só tenho a agradecer de fato.

À minha família, acima de tudo. Minha mãe, que sempre me deu exemplo de tranquilidade e sabedoria. Meu pai, que sempre me inspirou perseverança e força de vontade. E minha irmã, companheira nas madrugadas de estudo.

“Se fechares a porta a todos os erros, a verdade ficará lá fora.”

(Rabindranath Tagore)

RESUMO

O presente estudo tem como objetivo a aplicação de cinco técnicas de classificação distintas para a construção de modelos de previsão de insolvência. As técnicas utilizadas foram Análise Discriminante Linear, Regressão Logística, Vizinhos Mais Próximos, Árvores de Classificação e Redes Neurais Artificiais. Para esse fim, foram utilizada informações contábeis de empresas brasileiras de capital aberto, compondo uma amostra com 87 companhias. Como objetivos secundários, essa pesquisa buscou comparar os resultados obtidos pelos métodos para indicar quais obtiveram melhor desempenho, e também quais indicadores foram os mais importantes para as análises. Os resultados obtidos para os cinco modelos indicam que é possível identificar com boa margem de certeza quais empresas se tornarão insolventes. O modelo com melhor performance foi o de Redes Neurais Artificiais. Os indicadores mais importantes foram o Endividamento Geral, a Margem Líquida e a relação entre o Capital Circulante Líquido e o Ativo Total.

Palavras-chave: Modelos de classificação, previsão de insolvência, empresas de capital aberto, indicadores contábeis.

ABSTRACT

This study aims to apply five different classification techniques to build insolvency prediction models. The methods used were Linear Discriminant Analysis, Logistic Regression, *k*-Nearest Neighbors, Classification Trees and Artificial Neural Networks. To this end, it was used accounting-based information of Brazilian joint-stock companies, which compounded a sample of 87 firms. As secondary objectives, this research compared the results that were obtained by each method and then tried to indicate which method performed better, and also which indicators were the most useful to the analysis. The results obtained to the five models indicate that it is possible to identify with good confidence, which companies are becoming insolvent. The method that achieved the best performance was the Artificial Neural Network. The most important indicators were the Equity to Total Liabilities, Net Profit Margin and the Working Capital to Total Assets.

Key-words: Classification models, insolvency prediction, Joint-stock companies, account-based indicators.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TALBEAS

SUMÁRIO

1. INTRODUÇÃO

A previsão de insolvência constitui uma ferramenta importante para o desenvolvimento econômico. Gestores de uma firma, por exemplo, podem usar esse tipo de informação para mudar decisões acerca do futuro de sua empresa, já que o processo de deterioração financeira ocorre, em geral, de maneira progressiva.

Uma das aplicações mais conhecidas da previsão de insolvência é auxiliar na redução do risco de crédito. Segundo Caouette *et al* (2008), sempre que um indivíduo faz uso de um produto ou serviço sem que haja o pagamento imediato pelo mesmo, é possível identificar elementos relacionados ao risco de crédito. Quando uma compra a prazo é realizada, a empresa que cede o produto está sendo exposta a um risco de não receber o valor devido. Ou ainda, empresas que fornecem serviços básicos de maneira contínua, como distribuição de água e energia elétrica, também se expõem ao risco de crédito, já que o pagamento pelos serviços prestados só é efetuado ao final de um certo período, normalmente mensal. Assim, o risco de crédito é, em outras palavras, a possibilidade de que um credor sofra uma perda em razão do não cumprimento de obrigações assumidas por terceiros.

No mercado financeiro, muitas transações envolvem a disponibilidade de recursos para tomadores, seja em forma de empréstimo ou financiamento, mediante o comprometimento de pagamento em uma data posterior. Com isso, é natural que bancos e outras instituições que forneçam crédito, ou mesmo investidores, busquem formas de otimizar a utilidade dos recursos repassados às empresas tomadoras de recursos, reduzindo ao máximo o risco de crédito.

Nesse sentido, a capacidade de prever insolvência tem um papel fundamental. Do ponto de vista econômico, diz-se que uma companhia está insolvente quando o total de seu passivo é superior ao seu ativo, ou seja, quando ela não pode pagar todas as dívidas assumidas mesmo com a liquidez total de seus bens e direitos. Por isso, estabelecer com antecedência quais empresas apresentam propensão à insolvência e quais são financeiramente saudáveis, é de vital importância para credores e investidores. Para ilustrar isso, analisa-se o caso das Lojas Arapuã, grupo empresarial que durante a década de 90 chegou a ser o maior varejista de eletrodomésticos no Brasil. Foi eleita na edição de julho de 1997 da revista Exame Melhores e Maiores como a melhor empresa de comércio varejista do ano anterior. No entanto, já em 1998 as Lojas Arapuã viram-se forçadas a pedir concordata por conta de dificuldades financeiras, o

que gerou um imbróglio judicial entre a empresa e parte de seus credores e fornecedores, ainda não resolvido até hoje.

Prever se uma empresa passará por dificuldades financeiras dentro dos próximos anos pode relevar-se uma tarefa muito difícil, muitas vezes impossível até, já que os problemas podem ser resultado de fatores fora do controle dos gestores, ligados a incertezas de mercado, por exemplo. Entretanto, é possível identificar, por meio de modelos de classificação, quais as organizações que possuem maior probabilidade de enfrentarem dificuldades em um futuro próximo.

O primeiro trabalho acadêmico feito com o objetivo de prever insolvência foi realizado por FitzPatrick em 1932. Intitulado *A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies*, comparava dados de 20 empresas falidas com uma mesma quantidade de não-falidas, emparelhadas por setor. Posteriormente, Beaver (1966) desenvolveu o primeiro modelo de previsão de insolvência baseado em indicadores financeiros, utilizando análise univariada. O estudo de Altman (1968) marca o início da utilização de técnicas de análise multivariada para a previsão de dificuldades financeiras. O autor valeu-se da Análise Discriminante para criar um modelo capaz de distinguir empresas solventes e insolventes, com base em seus indicadores contábeis de um ano após o evento. O estudo de Altman representou um salto qualitativo muito grande, o que fez com que o seu trabalho virasse referência para os estudos da área até aos dias atuais.

Desde o primeiro modelo proposto por Altman, muitos outros foram surgindo, e conforme a tecnologia foi avançando, novas técnicas mais poderosas puderam ser aplicadas ao problema de se prever a entrada em estado de insolvência. Atualmente é possível empregar, além das técnicas estatísticas clássicas como a análise discriminante e a regressão logística, técnicas de inteligência computacional, como é o caso das redes neurais artificiais e máquinas de suporte vetorial.

No Brasil, os estudos de previsão de insolvência tiveram seus primeiros passos com o trabalho de Kantiz (1976), que utilizou análise de balanços para identificar variáveis com elevado poder de discriminar empresas solventes de insolventes. O primeiro modelo brasileiro que fez uso de uma técnica multivariada foi o de Elizabetsky (1976), que utilizou a análise discriminante para classificar clientes do Banco Comercial. Estudos com técnicas mais sofisticadas ainda não são tão frequentes no Brasil, porém é possível identificar alguns trabalhos

que fizeram uso de técnicas como as redes neurais, análise por envoltória de dados, entre outras técnicas.

O objetivo deste trabalho é criar modelos de previsão de insolvência para empresas de capital aberto, cujas ações tenham sido negociadas na Bolsa de Valores de São Paulo, entre os anos de 2003 e 2012, com base em indicadores contábeis obtidos por meio do *software Economatica* e utilizando técnicas de classificação variadas. Como objetivos secundários, busca-se verificar quais indicadores possuem maior eficiência para a problemática da previsão de insolvência, e também quais técnicas estatísticas apresentam melhor desempenho.

O evento estudado foi a entrada em estado de insolvência, e o seu início foi definido como a data em que uma empresa tenha feito pedido formal de concordata ou recuperação judicial. A amostra final contou com 21 empresas insolventes, e 66 empresas solventes escolhidas conforme a distribuição setorial do primeiro grupo. As técnicas utilizadas para a construção dos modelos foram a análise discriminante, a regressão logística, os vizinhos mais próximos, as árvores de classificação e as redes neurais. O *software* usado foi o R: *A language and enviroment for statistical computing* (R CORE TEAM, 2013).

Este trabalho contém cinco capítulos, incluindo esta introdução. No capítulo 2 é feita uma revisão dos estudos anteriores acerca do tema abordado. No terceiro capítulo é tratada a metodologia utilizada na pesquisa e elaboração do trabalho. No quarto capítulo é feita uma análise dos resultados obtidos. No último capítulo são feitas considerações finais sobre o que foi exposto.

2. REVISÃO DA LITERATURA

Neste capítulo é feita uma abordagem inicial sobre a previsão de insolvência, iniciando-se com algumas definições sobre o termo em seu aspecto econômico e financeiro. Tratam-se também dos indicadores contábeis como variáveis preditoras de insolvência. Por fim, analisam-se na seção 2.3 alguns estudos importantes sobre o assunto, tanto no mundo como no Brasil, levando em conta a técnica estatística utilizada para a elaboração.

2.1 Insolvência de empresas

Insolvência, do ponto de vista empresarial, pode ser entendida de diversas maneiras. Chung, Tan e Holdsworth (2008) citam dois conceitos para insolvência. O primeiro deles está ligado ao fluxo de caixa e diz que uma companhia que não pode pagar suas dívidas antes do vencimento, é considerada insolvente. Esta situação, no entanto, pode ser fruto da incapacidade da empresa de realizar seus ativos antes dos vencimentos dos débitos, e não necessariamente da insuficiência deles. Isso abre margem para outra definição de insolvência dada pelos autores, ligada ao balanço, na qual uma empresa é considerada insolvente se o total do passivo exceder o total do ativo, mesmo que a companhia possa honrar os seus compromissos de curto prazo.

Para Altman (1968), a insolvência de uma empresa pode ser percebida quando suas ações retornarem menores dividendos aos seus investidores do que aqueles de outros ativos financeiros de risco semelhante disponíveis no mercado.

Segundo Ross (*apud* SILVA, 2006) o conceito de insolvência não é definido de maneira precisa. Há uma grande variedade de eventos que podem caracterizar, seja de maneira isolada ou em conjunto, o estado de insolvência em uma empresa. O mesmo autor ainda lista alguns sintomas da insolvência: i) redução de dividendos; ii) fechamento de unidades; iii) prejuízos; iv) dispensa de funcionários; v) renúncias de presidentes; vi) quedas substanciais do preço da ação.

Altman (1968) considerou insolventes em seu estudo, as empresas que fizeram pedido de concordata segundo a legislação americana. A Constituição dos Estados Unidos define como insolvente uma entidade cujo total de dívidas exceda a soma de seus direitos,

considerados em valor justo, ou que não tenham capacidade de quitar as suas obrigações, na medida em que elas passem a ser exigíveis.

No Brasil, o primeiro dispositivo legal que criava instrumentos jurídicos para o enfrentamento de dificuldades financeiras de empresas foi o Decreto Lei nº 7.665/1945, a Lei de Falências e Concordatas. A concordata era definida como uma ação na qual uma empresa devedora poderia renegociar os prazos de vencimento de dívidas ou reemitir débitos, objetivando a solução do seu passivo quirografário, e consequentemente, evitar ou suspender o processo de falência, em caso de concordata preventiva ou suspensiva.

Em 2005, com o surgimento da Lei nº 11.101, Lei de Recuperação de Empresas e Falências, a concordata foi substituída pela recuperação judicial. A concordata definida pela lei anterior possuía uma conotação de favorecimento legal, uma vez que era concedida por um juiz que julgasse que a empresa devedora agia de boa-fé, independente da concordância ou não dos credores. A recuperação judicial assume um caráter contratual, já que para que possa ser efetivada é necessário que 3/5 dos credores a aprovem, o que torna o seu cumprimento obrigatório para todas as partes. A recuperação judicial só pode ser iniciada antes do processo de falência, diferente da concordata que também poderia ser iniciada durante a falência, tendo efeito suspensivo. A recuperação judicial não pode ser requerida por empresas públicas, sociedades de economia mista, instituições financeiras, cooperativas de crédito, consórcios, entidades de previdência complementar, planos assistenciais de saúde, sociedades securitárias e sociedades de capitalização (CLARO, 2008).

A nova lei também mudou o entendimento sobre falência. Antigamente a sua finalidade era fazer com que uma empresa pagasse o que era devido aos seus credores, ou o que fosse possível ser liquidado. Com a mudança, a falência passou a ser vista como o processo de retirada de empresas irrecuperáveis do mercado. Somente após a retirada da empresa é que se preocupará com a quitação das dívidas, o que passou a denominar-se judicialmente de liquidação.

Estudos de previsão de insolvência que utilizam empresas brasileiras costumeiramente definem como evento a entrada em concordata, recuperação judicial ou falência. Isso pode ser visto, por exemplo, em Altman, Baydia e Dias (1979) e Horta (2010). Essa abordagem é a mais comum nesse tipo de pesquisa.

Existem também aqueles que, por utilizarem uma base de dados de bancos ou outras instituições interessadas nesse tipo de estudo, utilizam uma classificação prévia de empresas

solventes e insolventes realizadas por essas instituições, como é o caso, por exemplo, do trabalho de Elizabetsky (1976), que utilizou dados de empresas clientes do Banco Comercial. Na bibliografia pesquisada ainda encontraram-se estudos publicados no Brasil que consideram como data de entrada em insolvência o dia em que as ações de determinada empresa passaram a ser negociadas como concordatárias na Bolsa de Valores de São Paulo, como é o caso de Sanvicente e Minardi (1998).

2.2 Variáveis preditoras de insolvência

Na modelagem de previsão de insolvência, as variáveis utilizadas podem ter origem em registros contábeis ou em informações do mercado. Vários estudos, no entanto, apontam que modelos baseados em dados contábeis são mais eficientes do que aqueles elaborados com base em dados do mercado, como pode ser visto em Reisz e Perlich (2007) e Agarwal e Taffler (2008). As variáveis originadas do mercado também são criticadas devido ao fato de ser necessário assumir a hipótese de eficiência do mercado.

Diante disso, e também por compor a maior parte da literatura de previsão de insolvência, o presente trabalho focou na utilização de indicadores originados de demonstrações contábeis para a construção dos modelos.

Quanto aos indicadores contábeis, os que aparecem com mais frequência em estudos desse tipo são aqueles ligados à liquidez, endividamento e rentabilidade. Beaver (1966) e Altman (1968) apontaram que a variável mais significante para previsão de insolvência relacionava o Capital de Giro com o Ativo Total. Em seu trabalho, Kanitz (1978) também apontou para índices de liquidez como sendo os mais adequados para a modelagem, porém utilizou os índices de liquidez geral, corrente e seca.

Sanvicente e Minardi (1998) realizaram um estudo com o objetivo de apontar quais indicadores seriam os mais significativos para previsão de concordatas. Avaliaram 14 indicadores, sendo cinco deles os mesmos utilizados por Altman (1968), e chegaram à conclusão de que os índices de liquidez eram os mais indicados, seguidos por índices de endividamento e rentabilidade, que também possuíam importância segundo os autores. O

indicador com maior poder de predição corroborou com o resultado encontrado por Altman (1968)

Bellovary, Giacomino e Akers (2007), que fizeram uma revisão dos principais estudos publicados na área de previsão de insolvência entre 1930 e 2004, em países da América do Norte, Europa, Ásia e na Austrália, apontaram que a variável mais utilizada é a de Retorno sobre o Ativo (ROA), presente em 54 estudos, seguido da Liquidez Corrente, presente 51 vezes, e do Capital de Giro sobre Ativo Total, utilizado 45 vezes. Os autores analisaram em sua pesquisa 165 artigos, desconsiderando aqueles que replicavam modelos já construídos anteriormente.

Para avaliar quais indicadores são os mais frequentemente utilizados em estudos brasileiros, foram analisados 23 trabalhos feitos entre os anos de 1976 e 2011. Constatou-se que índices de liquidez são os mais recorrentes, em especial, a relação entre o capital circulante líquido com o ativo total. Índices de endividamento e rentabilidade também são bastante comuns, embora um pouco menos que os de liquidez. O índice de endividamento geral e a rentabilidade sobre os investimentos totais, muitas vezes chamada de rentabilidade sobre o ativo, foram os índices mais frequentes nos grupos de endividamento e rentabilidade, respectivamente. Há ainda estudos que utilizam índices de atividade e outros que normalmente não são classificados em nenhum grupo, estes, no entanto aparecem com bem menor frequência.

2.3 Modelos de previsão de insolvência

O primeiro trabalho encontrado na literatura pesquisada que tratou de índices contábeis e da sua relação com a insolvência de empresas foi o *A Test Analysis of Unsuccessful Industrial Companies*, boletim publicado pelo *Bureau of Business Research* em 1930, no qual foram analisados 24 indicadores de 29 empresas industriais em processo de falência. A média de cada um dos indicadores de todas as empresas foi aferida e depois comparada a cada indicador de cada firma individualmente, sendo notada uma similaridade entre o comportamento das empresas.

Em 1932, Fitzpatrick escreveu o artigo *A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies*, no qual comparou 13 indicadores de

19 empresas falidas, com os de 19 não-falidas, emparelhadas de acordo com o setor econômico. Chegou à conclusão de que, na maioria dos casos, empresas bem sucedidas apresentavam indicadores superiores aos de empresas que vieram a falir. Fitzpatrick (*apud* Anjum, 2012) identificou cinco estágios que antecedem a falência: i) incubação, quando os problemas financeiros começam a surgir; ii) constrangimento financeiro, estágio no qual a administração toma ciência das dificuldades da empresa; iii) insolvência financeira, quando a firma é incapaz de adquirir recursos suficientes para cobrir suas obrigações; iv) insolvência geral, que ocorre quando o total do passivo excede o ativo; e v) insolvência legal, quando há procedimentos legais que buscam proteger os credores da empresa ou quando ocorre a sua liquidação.

Em 1966, Beaver levou o estudo da previsão de insolvência para um novo estágio. Em seu trabalho *Financial Ratios as Predictors of Failure*, Beaver utilizou os dados de empresas industriais dos EUA que faliram entre os anos de 1954 e 1964, emparelhadas, por setor e tamanho do ativo, com empresas economicamente saudáveis do mesmo período. Foram considerados no estudo os cinco anos que antecederam o pedido formal de falência. A amostra inicial contava com 158 empresas, porém, devido à indisponibilidade dos dados ao longo dos cinco anos, essa quantidade foi diminuindo, chegando a 117 no quinto ano que antecedeu a falência. Com base nas demonstrações financeiras destas empresas, Beaver montou 30 índices, divididos em seis grupos. Posteriormente escolheu um índice de cada grupo, conforme mostra o Quadro 1.

Quadro 1 – Variáveis utilizadas por Beaver (1966)

Índice 1:	$\frac{\text{Fluxo de Caixa}}{\text{Total das Dívidas}}$	Índice 4:	$\frac{\text{Capital de Giro}}{\text{Ativo Total}}$
Índice 2:	$\frac{\text{Lucro Líquido}}{\text{Ativo Total}}$	Índice 5:	$\frac{\text{Ativo Circulante}}{\text{Passivo Circulante}}$
Índice 3:	$\frac{\text{Total das Dívidas}}{\text{Ativo Total}}$	Índice 6:	$\frac{\text{Ativo Não Operacional – Passivo Circulante}}{\text{Despesas Operacionais}}$

Fonte: Beaver, (1966 apud Silva, 2006)

Utilizando a média das empresas solventes e insolventes em cada um desses indicadores ao longo dos cinco anos, Beaver fez, inicialmente, uma análise de perfil e concluiu que os índices das empresas falidas se deterioraram com muito mais rapidez do que os das empresas que permaneceram saudáveis. Posteriormente, o autor testou a habilidade preditiva de cada um dos indicadores, montando modelos com cada um deles. Ao concluir o seu trabalho, Beaver sugere que estudos posteriores utilizem vários indicadores simultaneamente na

construção dos modelos, o que acabaria por determinar a tendência dos trabalhos vindouros acerca da previsão de insolvência.

No Brasil, Kanitz desenvolveu um estudo univariado em 1976. Considerava que os balanços patrimoniais das empresas brasileiras, especialmente de empresas de pequeno e médio porte, apresentavam valores que não representavam bem sua realidade, além de despadronizados, prejudicando a qualidade da análise. Apesar das críticas, Kanitz acreditava que um balanço bem analisado poderia ainda fornecer boas informações quando confrontados com os de outras empresas. Em seu estudo, utilizou 516 índices de 21 empresas insolventes entre os anos de 1972 e 1974 emparelhadas com outras 21 empresas solventes, tendo como critérios o setor de atuação e o porte. Os demonstrativos utilizados foram de dois anos anteriores à falência. Para todas as empresas foi determinado seu percentil em cada um dos indicadores, ou seja, o percentual de empresas que possuíam desempenho pior que o de determinada empresa, ou seja, uma empresa situada no percentil 70 possuía um indicador melhor do que 70% das demais. Kanitz então buscou quais dos indicadores listavam como piores as empresas que efetivamente eram insolventes, chegando à conclusão de que 81 indicadores eram significantes para a previsão de insolvência, e que a deficiência dos balanços das empresas brasileiras não afeta este tipo de análise, muito embora fosse desejável a correção das imperfeições para um resultado mais preciso.

As próximas seções tratarão dos estudos feitos utilizando modelos de análise multivariada, segregando-os segundo as técnicas utilizadas, conforme estas foram surgindo ao longo dos anos.

2.3.1 Análise Discriminante

O primeiro estudo que utilizou alguma forma de análise multivariada para fins de previsão de insolvência foi o de Altman (1968). A técnica escolhida pelo autor foi a Análise Discriminante Linear (LDA). Naquela época havia uma predominância na área financeira da Análise de Regressão Múltipla, porém, Altman julgou a LDA mais apropriada para o estudo ao qual se propusera.

O modelo de LDA objetiva classificar uma observação dentro de um ou mais grupos levando em consideração as suas características individuais. É indicada quando a variável dependente é qualitativa (solvente e insolvente para os estudos de previsão de insolvência), enquanto a Regressão Múltipla usa variáveis dependentes quantitativas. Para que essa técnica seja aplicada é necessário que os grupos sejam estabelecidos, para a partir daí, com base nas características de cada grupo, ou seja, as variáveis independentes (índices financeiros), a LDA extraia uma combinação linear desses fatores que seja capaz de discriminá-los dentro dos grupos previamente estabelecidos. Essa técnica surgiu, segundo Corrar *et al* (2012), em 1935, em um estudo para classificação de plantas em duas populações feito por R. A. Fisher.

A função discriminante obtida apresenta a forma $Z = v_1X_1 + v_2X_2 + ... + v_nX_n$, em que $v_1, v_2, ..., v_n$ são os coeficientes de discriminação e $X_1, X_2, ..., X_n$ são as variáveis independentes. Para cada variável dependente, é calculado um escore discriminante Z, utilizado na classificação.

Este método tem como premissas a normalidade e a homocedasticidade das variáveis independentes dentro dos grupos que compoem a amostra, ou seja, os indicadores contábeis do grupo de empresas solventes e insolventes devem seguir uma distribuição normal e terem matrizes de variância-covariância semelhantes. Vale ressaltar que a normalidade e a homocedasticidade que devem ser testadas é a multivariada, ou seja, considerando coletivamente todas as variáveis utilizadas na amostra. É comum que trabalhos de previsão de insolvência limitem-se a testar a normalidade e a homocedasticidade de maneira univariada, porém, mesmo que todas as variáveis atendam isoladamente a esses pressupostos, isso não garante que o vetor das variáveis consideradas como um todo também atenda.

Em seu estudo, Altman (1968) utilizou uma amostra de 66 empresas, 33 em cada grupo. As empresas consideradas insolventes no estudo foram aquelas que fizeram pedido judicial de concordata de acordo com a lei americana, entre os anos de 1946 e 1965, todas do setor industrial. As empresas solventes foram escolhidas de maneira pareada às insolventes, levando em conta o ramo de atuação, e tamanho do ativo que não extrapolasse os limites do primeiro grupo, que ficara $ 0,7 milhão e $ 25,9 milhões.

Altman selecionou, inicialmente, 22 variáveis, classificadas em cinco grupos: liquidez, lucratividade, alavancagem, solvência e atividade. Algumas delas foram escolhidas com base na frequência com que elas foram citadas na literatura e na sua potencial relevância para o estudo. Outras foram adicionadas pelo autor em caráter experimental. Altman esperava

que o seu modelo final possuísse poucas variáveis já que muitas das que foram selecionadas possuíam forte correlação, podendo assim reduzir a sua quantidade. Isso justifica-se, pois uma variável que é fortemente correlacionada com outra do modelo.

Dentre as 22 variáveis, foram escolhidas cinco que obtiveram o melhor resultado juntas na previsão de insolvência. Para chegar a essa conclusão, Altman realizou os seguintes procedimentos: i) observação da significância estatística de várias funções alternativas, incluindo a determinação da contribuição relativa de cada variável independente; ii) avaliação da intercorrelação entre as variáveis relevantes; iii) observação da acurácia da previsão dos vários perfis; e iv) julgamento do analista. O modelo final e as variáveis que o compõem estão listadas no Quadro 2.

Quadro 2 – Fórmula discriminante de Altman (1968)

$Z = 0{,}012X_1 + 0{,}014X_2 + 0{,}033X_3 + 0{,}006X_4 + 0{,}999X_5$
Z = Índice geral
X_1 = Capital de Giro / Ativo Total
X_2 = Lucros Retidos / Ativo Total
X_3 = Lucros Antes dos Juros e Imposto de Renda (LAJIR) / Ativo Total
X_4 = Valor de mercado do Patrimônio Líquido / Valor Contábil do Passivo Total
X_5 = Vendas / Ativo Total

Fonte: Altman (1968).

Os resultados obtidos na classificação das 66 empresas que compuseram a amostra foram resumidos nas Tabelas 1 e 2, para um e dois anos antes do pedido de concordata, respectivamente.

Tabela 1 – Classificação para o ano anterior à concordata

	Classificadas como:		
	Insolventes	Solventes	Total
Insolventes	31 93,94%	2 6,06%	33 100,00%
Solventes	1 3,03%	32 96,97%	33 100,00%
Total	32	34	66

Fonte: Altman (1968).

Tabela 2 – Classificação para dois anos anteriores à concordata

	Classificadas como:		
	Insolventes	Solventes	Total
Insolventes	23	9	32
	71,88%	28,13%	100,00%
Solventes	2	32	33
	6,06%	93.94%	100,00%
Total	25	40	65

Fonte: Altman (1968).

O modelo obteve uma precisão geral de 95,45% para um ano antes do evento da concordata e 83,08% para dois anos. Altman (1968) também faz uma diferenciação entre os tipos de erro do modelo. Ele chama de Erro Tipo I aquele em que uma empresa insolvente é classificada como solvente, enquanto o Erro Tipo II é a classificação de uma empresa solvente como insolvente. A literatura sugere que o primeiro tipo de erro é mais preocupante do que o segundo, uma vez que para um credor é mais prejudicial perder o investimento feito em uma empresa que venha a quebrar, do que o custo de oportunidade de deixar de investir em uma empresa saudável (CAOUETTE *apud* CASTRO JÚNIOR, 2003). Nesse modelo, o Erro Tipo I foi de 6,06% e 28,13% e o Erro Tipo II de 3,03% e 6,06%, para um e dois anos anteriores à falência, respectivamente.

Para validar os resultados, Altman (1968) realizou quatro testes diferentes. O primeiro foi feito com base em cinco sub-amostras compostas por 16 empresas sorteadas aleatoriamente. Os percentuais de classificação correta desse teste oscilaram entre 91,18% e 97,06%. Para o segundo teste, foi selecionada uma nova amostra de 25 empresas, toda insolventes, dentro dos mesmos setores e tamanho do ativo daquelas que compuseram a amostra original. Nesse teste, 24 empresas foram classificadas corretamente, ou seja, apresentou 96% de precisão, valor que superou inclusive o obtido com a amostra original.

No terceiro teste, Altman selecionou 66 empresas que demonstraram perdas em seus balanços entre os anos de 1958 e 1961, mas que não haviam feito pedido judicial de concordata. Nessa amostra, apenas a restrição setorial foi mantida. O modelo classificou 52 empresas como solventes e 14 como insolventes, representando um percentual de acerto de 78,79%, que mostrou-se bastante surpreendente, já que as empresas utilizadas nessa amostra, apesar de não terem feito pedido de concordada, tiveram um desempenho que pode ser considerado inferior ao esperado pelo mercado.

O último teste realizado pelo autor utilizou dados dos cinco anos que antecederam o pedido de concordada das empresas. A escolha desse período de tempo baseou-se em estudos como o de Beaver (1966), que indicava que uma tendência à insolvência poderia ser percebida a partir de cinco anos anteriores a sua consumação. O resultado obtido por Altman, no entanto, mostrou-se satisfatório apenas para dois anos, conforme mostra a Tabela 3.

Tabela 3 – Classificação das empresas utilizando cinco anos

Ano anterior à concordata	Quantidade de empresas	Acertos	Erros	Percentual de acerto
1º	33	31	2	93,94%
2º	32	23	9	71,88%
3º	29	14	15	48,28%
4º	28	8	20	28,57%
5º	25	9	16	36,00%

Fonte: Altman (1968).

Com base nos resultados, Altman (1968) foi capaz ainda de estabelecer pontos de corte segundo os valores do escore discriminante. Empresas cujos escores foram superiores a 2,99 foram todas classificadas como solventes, enquanto aquelas com escore menor que 1,81 caíram no grupo de insolvência. Á região entre 1,81 e 2,99, deu o nome de "zona de ignorância" ou "área cinza", devido à suscetibilidade de erro de classificação dos valores situados nela.

A pesquisa realizada por Altman (1968) representou um enorme avanço na área de previsão de insolvência. Como consequência disso, a Análise Discriminante é a técnica multivariada mais encontrada nos trabalhos publicados. Segundo Bellovary, Giacomino e Akers (2007), a técnica utilizada por Altman (1968) foi predominante até o final da década de 80, tendo sido até então contabilizados 78 modelos, dentre os quais 52 utilizaram a Análise Discriminante, ressaltando que dos 26 modelos que não a utilizaram, 20 foram elaborados durante a década de 80.

No Brasil, o primeiro estudo que teve como foco a previsão de insolvência utilizando a LDA foi feito por Elisabetsky (1976). Este utilizou a Análise Discriminante para desenvolver três modelos diferentes, sendo um com cinco variáveis, outro com dez e um último com quinze. A amostra do estudo era composta por 99 empresas insolventes do ramo de confecção, emparelhadas com outras 274 do mesmo ramo sem dificuldades financeiras. O período analisado foi de 1972 a 1975. Foram considerados inicialmente 60 índices contábeis,

porém índices altamente correlacionados foram excluídos, restando 38 que foram submetidos ao processo de *stepwise*, no qual as variáveis vão sendo incluídas no modelo até um ponto que a adição de uma nova variável não implique em melhoria na capacidade preditiva. A equação final possuía 28 variáveis que classificaram todas as empresas corretamente. No entanto, devido à dificuldade que a utilização de tantas variáveis poderia proporcionar, o autor montou três modelos com uma quantidade menor delas: o primeiro com cinco variáveis obteve um erro geral de 31,48%, 25,93% para erro tipo I e 37,04% de erro tipo II; o segundo modelo com 10 variáveis classificou erroneamente 18,52% das empresas, com 22,22% de erro do tipo I e 14,81% do tipo II; e por fim o modelo com 15 variáveis errou em 14,81% das empresas, com percentual de erro tipo I de 18,52% e 11,11% para o erro tipo II.

Kanitz (1978) construiu um modelo baseado em Análise Discriminante que ficou conhecido como Termômetro de Kanitz. A função discriminante do modelo, chamada pelo autor de Fator de Insolvência, possuía cinco índices, sendo três deles de liquidez (geral corrente e seca). Nesse modelo, Empresas com escore superior a zero são tidas como solventes, enquanto empresas com escore inferior a -3 classificam-se como insolvente. A região entre -3 e zero é chamada de Zona de Penumbra, e deve servir como sinal de alerta para os gestores.

Vários outros estudos importantes foram desenvolvidos no Brasil utilizando a LDA, entre eles Matias (1978), Altman, Baidya e Dias (1979), que adaptaram o modelo desenvolvido por Altman para empresas brasileiras, Silva (1982), Kasznar (1986), Carmo (1987) e Santos (1996).

2.3.2 Regressão Logística

Em estudos de previsão de insolvência, a variável resposta é dicotômica, ou seja, possui apenas duas possibilidades, solvente e insolvente. Em casos desse tipo a Regressão Logística (RL) mostra-se um modelo bastante apropriado. Segundo Hair *et al* (2005), esse tipo de regressão apresenta uma gama de vantagens sobre a Análise Discriminante, já que o primeiro é muito mais robusto quando às hipóteses de normalidade e homocedasticidade não são atendidas.

Outra vantagem da RL sobre a LDA é o fato de que o escore obtido pelo segundo modelo possui pouca interpretação intuitiva, enquanto o valor obtido pela RL pode ser associado à probabilidade de ocorrência do evento estudado.

Algumas considerações teóricas e práticas sugerem que, quando a variável resposta é binária, a forma da função resposta será frequentemente sigmoidal. Para esses casos a função logística é a mais utilizada. Sua expressão matemática é dada por:

$$E[Y] = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} = [1 + \exp(-\beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p)]^{-1}$$

em que Y é a variável resposta, X_1, ..., X_p são as variáveis explicativas e β_0, ..., β_p são os coeficientes estimados a partir do conjunto de dados por meio da máxima verossimilhança.

Uma propriedade bastante útil da função logística é o fato dela poder ser linearizada. Definindo $\pi(x) = E[Y|x] = P(Y=1|x)$, pode ser aplicada a transformação *logit*, dada por:

$$\text{logit}[\pi(x)] = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right)$$

logo, a função resposta *logit* pode ser expresso como:

$$\text{logit}[\pi(x)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

A razão entre $\pi(x)$ e $1 - \pi(x)$ é chamada de *odds ratio*, ou razão de chance, pois representa uma relação entre as probabilidades de sucesso e fracasso.

O modelo de regressão logística tem como base os seguintes pressupostos: i) a contribuição de cada variável X_p, p = 1, ..., n é proporcional ao seu coeficiente β_p; ii) a contribuição de uma variável independente é constante e não depende das outras variáveis independentes; iii) Os erros são independentes e apresentam distribuição binomial; iv) as variáveis independentes não são multicolineares; e v) a escala logit[$\pi(x)$] é aditiva e linear.

De acordo com Corrar *et al* (2012), a técnica foi desenvolvida durante a década de 1960, sendo o *Framingham Heart Study*, realizado pelo National Heart, Lung and Blood Institute – NHLBI e a Universidade de Boston, um dos primeiros estudos que mais contribuíram para a sua notoriedade. A RL foi utilizada para identificar fatores responsáveis pelo desencadeamento de doenças cardiovasculares.

Na área de previsão de insolvência, a técnica começou a ser empregada no final da década de 70 com o trabalho de Martin (1977), porém o estudo realizado por Ohlson (1980) é

o mais conhecido quando se trata de previsão de insolvência com uso da RL. Uma das grandes contribuições da pesquisa de Ohlson (1980) foi levar em consideração a data da publicação das demonstrações contábeis para a composição da base de dados. Os estudos feitos anteriormente não tiveram essa preocupação, no entanto, o autor argumentava que a capacidade de previsão dos modelos depende da disponibilidade das variáveis independentes para utilização antes que o evento a ser estudado ocorra. Ao se utilizarem demonstrativos publicados após o pedido de concordata, há uma sobrevalorização da capacidade de previsão do modelo encontrado. As empresas consideradas insolventes foram aquelas que fizeram pedido legal de concordata entre 1970 e 1976. Foram escolhidas 105 empresas com problemas financeiros e 2.058 saudáveis, com a exigência de que nenhuma delas fizesse parte do setor de serviços, transporte ou financeiro, e que tivessem ações negociadas pelo menos nos três anos anteriores ao pedido de concordata. Não houve restrições quanto ao tamanho do ativo ou porte da empresa. Ohlson (1980) construiu dois modelos, o primeiro apresentou 12,4% e 17,4% de erros do tipo I e II respectivamente, com ponto de corte 0,038. O segundo modelo classificou erradamente 20,2% das empresas insolventes e 8,6% das solventes, sendo seu ponto de corte igual a 0,08.

No Brasil, destaca-se o trabalho de Matias e Siqueira (1996) que utilizou a RL para construir um modelo de previsão de insolvência voltado para o setor bancário brasileiro. Os bancos considerados insolventes para o estudo foram aqueles em processo de liquidação ou intervenção iniciado entre julho de 1994 e março de 1995, sendo observados 16 casos nesse período. A amostra também foi composta por 20 bancos solventes durante o mesmo período, selecionados entre os que apresentaram melhor desempenho. O modelo final classificou corretamente 87,50% das empresas insolventes e 95,00% das solventes, e o ponto de corte considerado foi 0,5. Para validar o modelo, Matias e Siqueira (1996) selecionaram uma nova amostra composta por instituições consideradas por eles como tradicionais, isto é, com mais de dez anos de atuação no setor bancário brasileiro, com acionistas com mais de dez anos na própria instituição e que possuíssem diretores atuando também há mais de uma década na área bancária, totalizando 17 instituições. Utilizando o mesmo ponto de corte, o modelo classificou apenas uma dessas instituições como insolvente, o que representa 94% de precisão. Uma segunda validação foi feita utilizando informações de bancos que entraram em processo de liquidação ou intervenção entre abril de 1995 e o mesmo mês do ano seguinte, totalizando 13 instituições, das quais sete foram classificadas como insolventes, conferindo ao modelo uma precisão de 54%.

2.3.3 Classificador dos vizinhos mais próximos

O Classificador dos vizinhos mais próximos (kNN) é um dos métodos mais simples para a classificação de uma amostra. O modelo baseia-se na distância entre pares de observações. Nele há dois parâmetros a serem escolhidos: a função da distância a ser utilizada e o valor dos *k* vizinhos mais próximos.

No que diz respeito ao primeiro parâmetro, a distância euclidiana é a mais utilizada. Considerando $\boldsymbol{x_n} = (x_{n1}, x_{n2}..., x_{np})$ e $\boldsymbol{x_k} = (x_{k1}, x_{k2}, ..., x_{kp})$, duas observações com dimensão *p*, a distância euclidiana entre elas é dada por:

$$\mathrm{d}(\boldsymbol{x_n}, \boldsymbol{x_k}) = \|\boldsymbol{x_n} - \boldsymbol{x_k}\| = \sqrt{(\boldsymbol{x_n} - \boldsymbol{x_k})^2} = \sqrt{\sum_{i=1}^{p} (x_{ni} - x_{ki})^2}$$

Depois de computadas todas as distâncias de $\boldsymbol{x}$ para as demais observações, as *k* mais próximas, sendo este um número inteiro maior que ou igual a um, irão constituir a vizinhança da observação $\boldsymbol{x}$. A classificação do modelo para $\boldsymbol{x}$, então, será dada pela moda da classificação da vizinhança. Em caso de existência de mais de uma moda, Elkan (2011) diz não haver consenso sobre a melhor forma de lidar com essa questão, optando-se muitas vezes por uma escolha aleatória entre as modas.

Hastie *et al* (2009 *apud* Rebouças, 2011) afirma que apesar da simplicidade do método e da sua incapacidade de avaliar a natureza das relações entre as variáveis independentes e a dependente, o kNN apresenta-se como uma boa opção para fins preditivos.

Arieshanti *et al* (2013) comparou modelos de previsão de insolvência criados a partir do método kNN com outros construídos por Redes Neurais e Máquina de Suporte Vetorial. Em seu estudo, o autor utilizou a base de dados disponibilizada por Wieslaw em seu *website*, que é composta por 112 empresas insolventes e 128 solventes, considerando um período de dois a cinco anos anterior ao pedido de concordata. Os modelos baseados em metodologias de kNN obtiveram 77,50% e 75,42% de precisão, aqueles feitos por Redes Neurais acertaram em 74,50% e 71,00%, enquanto os baseados em Máquinas de Suporte Vetorial obtiveram 71,58% e 70,42%.

2.3.4 Árvores de classificação

Segundo Basgalupp (2010), o algoritmo de Árvores de Classificação e Regressão (*Classification and Regression Trees* – CART), foi proposto por Breiman *et al* (1984) e consiste em uma técnica não paramétrica que induz tanto árvores de classificação, caso a variável dependente seja categórica, quanto árvores de regressão, caso a variável dependente seja contínua. Ainda segundo o autor, uma das maiores virtudes da CART é a capacidade de pesquisa de relações entre os dados, mesmo que não sejam evidentes.

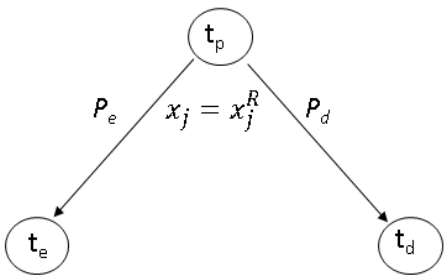
O método CART baseia-se na execução de partições binárias sucessivas de uma amostra, com base nos resultados amostrados das variáveis independentes, buscando a constituição de subamostras internamente homogêneas. A classificação dessas subamostras é realizada conforme alguma medida descritiva e a predição de novos elementos, executada por meio da estrutura de classificação constituída (TACONELLI, ZOCCHI e DIAS, 2009).

Os componentes elementares do modelo são os nós e as regras de divisão (*splitting rules*). O primeiro nó de uma árvore é chamado de raiz e representa todo o conjunto de dados. Os nós terminais recebem o nome de folhas. Os nós que dão origem a outros são chamados de pais, enquanto aqueles gerados são denominados de filhos.

Considere uma amostra cuja matriz de variáveis independentes $\boldsymbol{X}$ possui p variáveis x_j e n observações. Considere também que o vetor $\boldsymbol{Y}$ de variáveis dependentes é composto por n observações distribuídas entre k classes. Seja t_p um nó pai e t_d, t_e sejam nós filhos gerados a partir do primeiro, denominados nó direito e esquerdo, respectivamente.

A árvore de classificação é formada de acordo com as *splitting rules*, que dividem a amostra em partes menores que possuam máxima homogeneidade interna. A Figura 1 mostra uma representação gráfica do algoritmo de divisão da CART.

Figura 1 - Algoritmo de divisão de uma CART



Fonte: Timofeev (2004)

Em que x_j representa a variável dependente da observação j, x_j^R denota o valor da variável x_j que melhor divide a amostra, P_e e P_d são as probabilidades associadas aos nós esquerdo e direito respectivamente.

A homogeneidade máxima de todos os nós filhos (t_f) pode ser definida por uma medida de impureza $i(t)$. Como a impureza do nó pai é constante para qualquer possibilidade de divisão, a homogeneidade máxima dos nós filhos é equivalente à maximização da variação da medida de impureza $\Delta i(t)$:

$$\Delta i(t) = i(t_p) - E[i(t_f)] = i(t_p) - P_e i(t_e) - P_d i(t_d)$$

Assim, as observações são classificadas por meio do problema de maximização a seguir:

$$\underset{x_j \leq x_j^R, j=1,\ldots,M}{\arg\max} \left[i(t_p) - P_e i(t_e) - P_d i(t_d) \right]$$

A equação acima demonstra que o algoritmo de uma CART buscará entre todas variáveis da matriz $\boldsymbol{X}$ aquele valor que atenda à condição $x_j \leq x_j^R$ e que maximize a variação da medida de impureza.

Para a definição da medida de impureza $i(t)$ existem várias funções que podem ser utilizadas, porém a mais utilizada é o índice de Gini:

$$i(t) = \sum_{a \neq b} p(a|t) p(b|t)$$

em que a, b representam as k classes das variáveis dependentes e $p(a|t)$ é a probabilidade condicional de ocorrência da classe a dentro do nó t.

Aplicando o índice de impureza de Gini ao problema de maximização descrito anteriormente, chega-se ao seguinte resultado:

$$\underset{x_j \leq x_j^R, j=1,\ldots,M}{\arg\max} \left[-\sum_{n=1}^{K} p^2(n|t_p) + P_e \sum_{n=1}^{K} p^2(n|t_e) + P_d \sum_{n=1}^{K} p^2(n|t_d) \right]$$

O algoritmo de Gini irá procurar pela classe com o maior número de indivíduos dentro da amostra, isolando-a do restante dos dados.

Timofeev (2004) aponta algumas vantagens das CART. Quando o método é utilizado para classificação, o próprio algoritmo indica quais variáveis são mais importantes,

descartando aquelas menos significantes, o que é bastante útil quando não se tem conhecimento prévio de quais variáveis são mais relevantes para a classificação a ser realizada. Outra vantagem do método é o fato de ser invariável a transformações monótonas realizadas nas variáveis independentes, ou seja, o resultado final não será alterado mesmo que sejam aplicadas aos dados operações como logaritmo e radiciação. As CART também lidam com *outliers* de maneira muito mais robusta do que outros métodos, já que o método permite que esses tipos de observações sejam isolados em um nó à parte das demais observações. Esta é uma característica muito importante já que *outliers* costumam ter efeitos negativos sobre os resultados de modelos estatísticos.

Santos *et at* (2006), utilizaram as CART, entre outros modelos, para classificar 2.288 empresas que estiveram em funcionamento entre 1999 e 2003, todas situadas na região norte de Portugal. Destas, 325 haviam pedido concordata durante o período, enquanto as outras 1.963 permaneceram solventes. Construíram com esses dados quatro modelos utilizando árvores de classificação. Dois modelos consideravam apenas um ano anterior à entrada em insolvência, enquanto os outros dois consideravam toda a informação que precedia o evento, ou seja, três anos. Outro ponto para a diferenciação dos modelos foi a quantidade de variáveis utilizadas para a classificação: dois modelos consideravam todos os indicadores construídos pelos autores, totalizando 58; os outros utilizavam apenas 11 variáveis consideradas por eles como as mais importantes. Para todos os modelos o conjunto de dados foi dividido em duas partes de maneira aleatória, sendo uma subamostra usada para o treinamento da árvore, enquanto a outra servia para validá-la. As precisões de acerto de todos os modelos foram bastante elevadas e podem ser observadas na Tabela 4.

Tabela 4 - Resultados do estudo de Santos *et al* (2007)

	58 indicadores		11 indicadores	
	Um ano	Três anos	Um ano	Três anos
Insolventes	86%	96%	95%	95%
Solventes	99%	90%	95%	95%
Total	97%	92%	95%	95%

Fonte: Santos *et al* (2007)

Horta *et al* (2011), com o intuito de testar métodos diferentes para seleção de indicadores para serem utilizados em estudos de previsão de insolvência, utilizou árvores de classificação. A base de dados desse trabalho foi composta, inicialmente, por empresas listadas no Serasa e na Bovespa como concordatárias, em recuperação judicial ou falidas, durante o período de 2005 a 2007. Posteriormente, buscaram outras empresas saudáveis que atuassem no

mesmo setor, com tamanho do ativo semelhante e, quando possível, localizadas na mesma região das empresas do primeiro grupo. A amostra final apresenta 56 empresas insolventes e 112 solventes. Para validar os resultados, foi utilizada a validação cruzada, que consiste na divisão do conjunto original de dados em *k* subconjuntos menores, sendo estimado um modelo utilizando *k*-1 desses grupos, que é validado com o conjunto que ficou de fora da estimação. Horta *et al* (2011) utilizaram dez subconjuntos para a validação cruzada. Foram construídos três modelos utilizando metodologias diferentes para a seleção dos dados, com percentuais de acerto de 89,88%, 91,66% e 92,26%.

2.3.5 Redes neurais artificiais

As Redes Neurais Artificiais (*Artificial Neural Network* – ANN) são uma técnica de processamento de informação inspirada pelo sistema nervoso humano. O cérebro humano processa informações de maneira diferente quando comparado a um computador convencional. Segundo Hakin (2001), o cérebro pode ser considerado um sistema de processamento de informação extremamente complexo, não linear e paralelo, com capacidade de organizar seus componentes estruturais, os neurônios, para realizar atividades como reconhecimento de padrões, percepção e controle motor, executando-as de maneira muito mais eficaz que sistemas computacionais.

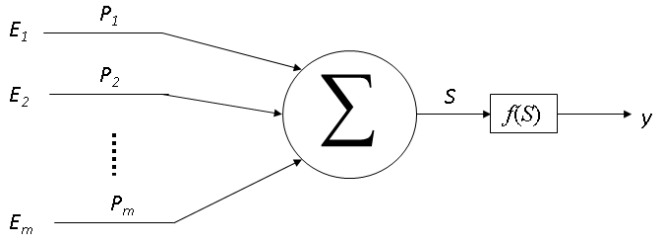
O primeiro modelo de ANN foi proposto por McCulloch e Pitts (1943), que propuseram um sistema para reproduzir as características básicas de um neurônio. O modelo McCulloch-Pitts é formado por uma série de entradas E_1, E_2, ..., E_n que dão origem a uma valor binário *y*:

$$S = \sum_{k=1}^{n} E_k P_k$$

$$y = f(S)$$

em que *f*(*S*) denota a função de ativação de Heaviside, assumindo o valor 1, caso S seja maior ou igual a zero, e valor nulo, caso S seja menor que zero. P_k representa os pesos associados às sinapses. Em caso de peso positivo, as sinapses são denominadas de excitação, caso o peso seja negativo, a sinapse é de inibição. A Figura 2 mostra uma ilustração do modelo McCulloch-Pitts.

Figura 2 - Representação de um modelo com um neurônio



Fonte: Kawaguchi (2000)

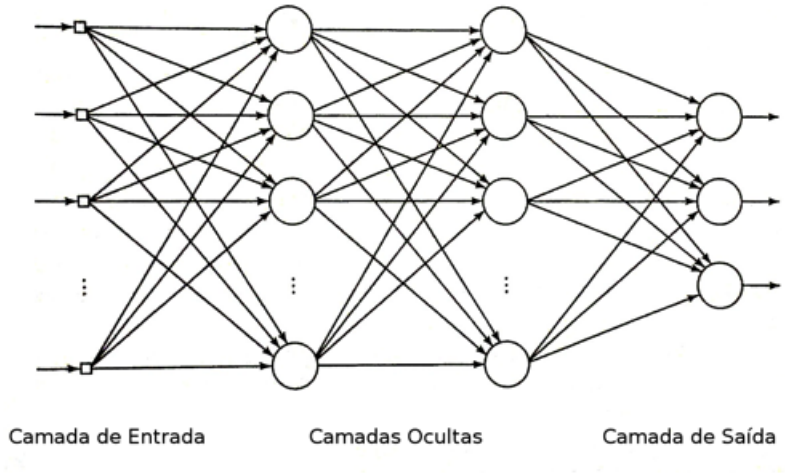
O grande avanço do modelo elaborado por McCulloch e Pitts foi demonstrar que elementos funcionais cujo funcionamento é bastante simples ganham capacidade de processamento e podem realizar tarefas muito mais complexas quando existe um sistema de conexões entre esses elementos.

Hebb (1949) propôs um postulado para a aprendizagem em nível celular no qual afirmava que dois neurônios que se encontram nas extremidades de uma sinapse, quando são ativados simultaneamente, o peso sináptico dessa ligação cresce de maneira seletiva. Stent (1973) e Changeux (1976) propuseram uma extensão ao postulado de Hebb, afirmando que caso dois neurônios nas duas extremidades da sinapse sejam acionados de maneira assíncrona, a força da sinapse que os une será enfraquecida ou eliminada. Ou seja, no que diz respeito a modelos do tipo ANN, uma sinapse Hebbiana é fortalecida quando os elementos pré e pós-sinápticos são positivamente correlacionados, e é enfraquecida quando os mesmos se correlacionam de maneira negativa.

Rosenblatt (1957) desenvolveu um modelo denominado perceptrão, que constitui a maneira mais simples de utilizar-se uma rede neural para a classificação de padrões linearmente separáveis (HAYKIN, 2001). O seu modelo era basicamente uma combinação do neurônio de McCulloch e Pitts com a regra de aprendizado de Hebb. O perceptrão de Rosenblatt é uma rede de camada única, na qual cada peso influencia uma única saída (REBOUÇAS, 2011).

Uma das limitações do perceptrão simples de Rosenblatt advém do fato de que este só é capaz de distinguir dados que sejam linearmente separáveis. Para superar isso é possível utilizar camadas escondidas, desde que disponham de uma quantidade adequada de unidades em cada uma delas. A essa rede constituída por mais de uma camada dá-se o nome de perceptrão de multicamada (*multilayer perceptron* – MLP), como mostra a Figura 3.

Figura 3 - Arquitetura de um MLP



Fonte: Haykin (2001)

Minsky e Papert (1969) fizeram duras críticas ao perceptrão de Rosenblatt e seus variantes, o que inclui os MLP. Segundo os autores, esses tipos de modelo eram incapazes de fazer generalizações globais baseadas em exemplos aprendidos localmente. Tal crítica fez nascer grande desconfiança nas redes neurais, o que acabou por desestimular as pesquisas para o desenvolvimento do método.

O esquecimento das redes neurais só teve seu fim em 1986 com a publicação do livro *Parallel Distributed Processing*, elaborado por McClelland e Rumelhart, no qual surgiu um método para o ajuste de parâmetros de redes não-recorrentes de múltiplas camadas que era baseado em um algoritmo de retropropagação (*backpropagation*). O algoritmo consiste em dois passos através das diferentes camadas da rede, sendo o primeiro um para frente, propagação, quando um padrão de atividade é aplicado aos nós sensoriais da rede e seu efeito se propaga através da rede. Durante esse passo os pesos sinápticos da rede são fixos. Já no segundo passo, dado para trás, os pesos sinápticos são ajustados de acordo com uma regra de correção de erro. A resposta real da rede é subtraída da resposta desejada produzindo um sinal de erro que é propagado de volta através da rede, daí o nome de retropropagação (HORTA, 2010).

Segundo Corrar *et al* (2012), as ANN têm sido muito aplicadas na área de negócios, com objetivos cada vez mais diversificados, e encontrando, muitas vezes, resultados superiores aos métodos estatísticos convencionalmente aplicados. A técnica tem complementado e enriquecido modelos estatísticos de inadimplência, riscos securitários e avaliação dos riscos associado aos papéis financeiros, entre outras aplicações.

As ANN possuem muitas vantagens quando comparada com outros métodos estatísticos. Elas são capazes de tratar dados qualitativos, não precisam atender pressupostos quanto às distribuições amostrais e são pouco sensíveis ao efeito provocado pelo seu tamanho. A multicolinearidade tem efeito menos consideráveis para as ANN, apesar de ainda assim ser recomendável eliminar variáveis altamente correlacionadas (CASTRO JÚNIOR, 2003).

Odom e Sharda (1990) elaboraram um estudo no qual utilizaram ANN baseada em MLP com retropropagação e LDA para previsão de insolvência. Tomaram como base o trabalho de Altman (1968), fazendo uso inclusive das mesmas variáveis. A amostra utilizada era composta por 128 empresas. Foram construídos três modelos, considerando subamostras de proporções diferentes entre as empresas solventes e insolventes para o treinamento dos modelos, 50/50, 80/20 e 90/10. Para a etapa de testes foram utilizadas 55 empresas, sendo 27 problemáticas e 28 saudáveis. Os resultados podem ser visualizados na Tabela 5.

Tabela 5 - Resultados do estudo de Odom e Sharda (1990)

Proporção:		50/50		20/80		10/90		
		Classificadas como:						
		Insolventes	Solventes	Insolventes	Solventes	Insolventes	Solventes	Total
ANN	Insolventes	22 81,48%	5 18,52%	21 77,78%	6 22,22%	21 77,78%	6 22,22%	27 100,00%
	Solventes	5 17,86%	23 82,14%	6 21,43%	22 78,57%	4 14,29%	24 85,71%	28 100,00%
LDA	Insolventes	16 59,26%	11 40,74%	19 70,37%	8 29,63%	16 59,26%	11 40,74%	27 100,00%
	Solventes	3 10,71%	25 89,29%	4 14,29%	24 85,71%	6 21,43%	22 78,57%	28 100,00%

Fonte: Odom e Sharda (1990, *apud* Castro Júnior, 2003).

É possível ver pelos resultados que para todas as proporções tomadas, o resultado das ANN superam os da LDA. Analisando as classificações feitas pelos modelos de maneira mais detalhada, os autores perceberam ainda que em todos os casos em que houve erro nas ANN, a LDA também errou.

Chung, Tan e Holdsworth (2008) realizaram um trabalho semelhante ao trabalho anteriormente citado de Odom e Sharda. Utilizaram os dados de 10 empresas com problemas financeiros, definido como pedido formal de concordata, detectados entre 2005 e 2007, e compararam com outras 35 empresas sem dificuldades no mesmo período, todas as empresas do setor financeiro da Nova Zelândia. Calcularam 36 indicadores utilizando balanços até três anos anteriores ao pedido de concordata. Compararam o desempenho obtido utilizando a LDA

e as ANN por meio de um teste t chegando à conclusão de que o segundo modelo é mais eficiente.

Lachtermacher e Espenchitt (2001) aplicaram as ANN e LDA para criação de modelos de previsão aplicados a empresas prestadoras de serviços à Petrobrás. A amostra do estudo continha 48 empresas com falência requeria ou decretada entre 1983 e 1993, e outras 35 empresas com bom desempenho durante o mesmo período, todas dos setores de construção civil, montagem industrial ou elaboração de projetos de engenharia. Neste estudo foram consideradas inicialmente 24 indicadores, porém, após análise de correlação, foram selecionados 10 índices. Os autores construíram um total de seis modelos, cinco deles utilizando redes neurais do tipo MLP, e apenas um com análise discriminante. A precisão de todos os modelos baseados em ANN superaram os resultados obtidos com a LDA.

A literatura sobre o assunto sugere que modelos de ANN tem, em geral, desempenho superior para a previsão de insolvência.

3. METODOLOGIA

Neste capítulo são detalhados os procedimentos adotados para a execução da pesquisa. Inicialmente classifica-se o tipo de pesquisa que foi realizado. Em seguida são descritas a população e os critérios de amostragem. Segue-se um levantamento dos procedimentos adotados para a coleta dos dados, bem como a descrição das variáveis utilizadas para o estudo. A seguir os passos empregados na aplicação dos métodos são elencados. Por fim é feita uma explicação das estratégias adotadas para a comparação do desempenho dos modelos.

3.1 Tipo de pesquisa

Conforme Lakatos e Markoni (2003) a pesquisa realizada nesse trabalho é do tipo documental, posto que a fonte de coleta dos dados está restrita a documentos. No caso deste estudo os dados utilizados foram coletados por meio do *software* Economática. Apesar da coleta ter sido feita com o auxílio de uma ferramenta paga, os dados obtidos são de domínio público, podendo ser acessados por meio do *site* da CVM ou BOVESPA.

Esta pesquisa é ainda bibliográfica, uma vez que nela foram analisadas produções literárias e acadêmicas realizadas anteriormente sobre a aplicação de técnicas de classificação ao problema da previsão de insolvência.

3.2 População e amostra

A população considerada nesse estudo são as empresas brasileiras de capital aberto que estiveram em atividade entre os anos de 2003 e 2012. A escolha dessa população dá-se pela relativa facilidade quanto à obtenção de dados para o estudo proposto, o que torna-a a população mais utilizada nesse tipo de pesquisa. Beaver (1966), ao utilizar dados de empresas industriais de capital aberto dos EUA, afirma que a sua escolha foi feita de maneira relutante, já que a probabilidade de entrada em insolvência nesse grupo é consideravelmente menor quando

comparada à de outros tipos de empresa, que são em geral, menores. O *software* Economática apresenta para o período de 2003 a 2012, informações de cerca de 640 empresas de capital aberto no Brasil.

Para a composição da amostra, inicialmente buscou-se no sítio da CVM, quais empresas haviam feito pedido de concordata preventiva ou petição inicial de recuperação judicial entre os anos de 2003 e 2012. Como resultado da busca, obteveram-se 22 empresas, sendo uma delas descartada por ter feito pedido de concordata preventiva em 1998, o qual foi aceito, no entanto a empresa não foi bem sucedida e em 2009, aproveitando-se da mudança na legislação brasileira ocorrida em 2005, fez um pedido de recuperação judicial.

As 21 empresas consideradas insolventes foram então classificadas de acordo com o segmento de atuação. Para este fim recorreu-se inicialmente aos sistemas de classificação existentes no Economática, que conta com quatro tipos de classificação diferente, um sistema próprio, o Setor Eco, com 20 categorias, e três sistemas baseados no sistema NAICS, o primeiro deles também com 20 categorias, o segundo com 92 e o terceiro com 313. Os níveis com classificação menos específicas foram descartados pois agrupavam muitas empresas com atividades pouco comparável. O nível mais detalhado também foi considerado inviável para o estudo pois a maioria das empresas possuía uma categoria própria. Adotou-se então o NAICS-2, que considera 92 tipos de setores. Ao ser efetuado o emparelhamento das empresas solventes com as insolventes, ainda assim existiram empresas que figuravam sozinhas em suas classificações, e também, devido à grande incidência de empresas de um mesmo setor no grupo de insolventes como pode ser visto na tabela 8, foi necessário complementar a classificação. Para isso, considerou-se a classificação por subsetor da BOVESPA, que conta com 42 categorias. O Apêndice A traz a relação completa das empresas com suas respectivas classificações setoriais consideradas.

A amostra final é composta por 87 empresas, sendo 21 insolventes e 66 solventes. O Ativo Total do primeiro grupo teve como extremos, em milhares, R$ 23.620 e R$ 12.935.080, enquanto os ativos do segundo grupo situaram-se entre R$ 68.216 e R$ 13.662.280. Cada empresa insolvente foi emparelhada com até no máximo cinco empresas solventes, por segmento e tamanho do Ativo Total, de acordo com a disponibilidade dos dados.

3.3 Coleta dos dados

Esse trabalho utilizou dados do Balanço Patrimonial e da Demonstração do Resultado do Exercício do ano anterior ao pedido inicial de concordata ou recuperação judicial das empresas insolventes, ou seja, são demonstrativos dos anos de 2002 a 2011. Para as empresas solventes, foram usadas informações do mesmo período das empresas do grupo de concordatárias com as quais estas se encontravam emparelhadas. Os dados foram obtidos com o auxílio da ferramenta Economática.

Com as informações obtidas, foram construídos 16 indicadores contábeis, cuja escolha foi baseada em estudos anteriores realizados no Brasil. A limitação aos estudos nacionais deve-se ao fato de que estes tendem a apontar variáveis preditoras de insolvência que sejam mais adequadas ao estudo, já que foram utilizadas empresas brasileiras. Pereira, Domínguez e Ocejo (2007) afirmam que a evidência empírica permite constatar que a escolha de indicadores que obtiveram bom desempenho em trabalhos prévios, leva, geralmente, a bons resultados. Os índices construídos também levaram em consideração a disponibilidade dos dados, já que algumas das informações retiradas do Economática não apresentavam valores para todas as empresas. Os indicadores construídos seguem a recomendação da literatura para estudos de insolvência, que indica a melhor adequação de índices de liquidez, rentabilidade e endividamento. Todos os indicadores construídos são descritos detalhadamente a seguir.

- X_1 – (Ativo Circulante – Passivo Circulante)/Ativo Total [AcPcAt]

Indicador encontrado nos estudos de Altman, Baydia e Dias (1979); Sanvicente e Minardi (1998); Lachtermacher e Espenchitt (2001); Minussi, Damascena e Ness Júnior (2002); Guimarães e Moreira (2008); Brito, Assaf Neto e Corrar (2009). É um índice que compara a diferença entre os recursos alocados em ativos de curto prazo e as obrigações de curto prazo em relação ao tamanho do ativo total. Quanto maior for este indicador, maior será a capacidade de pagamento de uma empresa no curto prazo. De acordo com Brito, Assaf Neto e Corrar (2009), pode ser considerado um indicador de análise dinâmica que avalia a situação financeira de uma empresa. Esse índice é frequentemente apontado como o mais relevante para estudos de previsão de insolvência.

- X_2 – LAJIR/Ativo Total [LajirAt]

Indicador encontrado nos estudos de Altman, Baydia e Dias (1979); Sanvicente e Minardi (1998). É um índice de rentabilidade que relaciona a lucratividade da empresa antes dos juros e Imposto de Renta (LAJIR) com o investimento total realizado, levando em consideração as estratégias operacionais adotadas, mas excluindo a depreciação e amortização, que não demandam contrapartida monetária imediata (HORTA, 2010).

- X_3 – Patrimônio Líquido/Exigível Total [PlExgt]

Indicador presente em Kanitz (1978); Altman, Baydia e Dias (1979); Sanvicente e Minardi (1998); Castro Júnior (2003); Onusic et al (2006); Guimarães e Moreira (2008); Horta (2010). Índice de endividamento que mostra a dependência da empresa em relação a recursos externos. Segundo Guimarães e Moreira (2008), o patrimônio líquido é, em última instância, a garantia para liquidação dos compromissos com terceiros, por isso, quanto menor o resultado desse índice, maior o risco do negócio.

- X_4 – Receita Líquida/Ativo Total [RlAt]

Indicador presente em Altman, Baydia e Dias (1979); Castro Júnior (2003); Onusic et al (2006); Brito, Assaf Neto e Corrar (2009); Horta (2010). É um índice de rentabilidade que verifica se o volume de vendas do período foi adequado ao capital total investido na empresa. Indica o nível de eficiência com o qual os recursos de uma empresa são investidos. Denominado de Giro de Ativo.

- X_5 – Exigível Total/Ativo Total [ExgtAt]

Indicador encontrado em Lachtermacher e Espenchitt (2001); Brito, Assaf Neto e Corrar (2009); Horta (2010). Índice de endividamento que avalia a proporção dos ativos totais de uma empresa financiados por credores. Denominado de Índice de Endividamento Geral.

- X_6 – Ativo Circulante/Passivo Circulante [AcPc]

Indicador presente em Matias (1976); Kanitz (1978); Castro Júnior (2003); Onusic (2006); Brito, Assaf Neto e Corrar (2009); Horta (2010). Índice de liquidez que demonstra a capacidade que uma empresa possui de pagar suas dívidas de curto prazo. Denominado de Índice de Liquidez Corrente.

- X_7 – Lucro Líquido/Ativo Total [LlAt]

Indicador presente em Lachtermacher e Espenchitt (2001); Castro Júnior (2003); Scarpel (2008); Horta (2010). Índice de rentabilidade que mostra o retorno total dos

investimentos feitos pela empresa. Representa a capacidade que os ativos têm de gerar lucros. Denominado de Retorno sobre o Investimento Total, ou Retorno sobre o Ativo.

- X_8 – (Ativo Circulante – Estoques)/Passivo Circulante [AcEstPc]

Indicador presente em Kantiz (1978); Onusic et al (2006); Castro Júnior (2003); Horta (2010). Índice de liquidez que indica a capacidade da empresa de pagar suas dívidas de curto prazo considerando apenas seus ativos mais líquidos. Pode ser entendido como uma versão mais conservadora do Índice de Liquidez Seca. Denominado de Índice de Liquidez Seca.

- X_9 – (Ativo Circulante + Realizável a Longo Prazo)/Exigível Total [AcRlpExgt]

Indicador presente em Kanitz (1978); Onusic (2006); Scarpel (2008); Horta (2010). Índice de liquidez que indica a capacidade que uma empresa possui de pagar todas as suas obrigações, sejam elas de curto ou longo prazo. Denominado de Índice de Liquidez Geral.

- X_{10} – Lajir/Despeas Financeiras [LajirDespfin]

Indicador encontrado em Sanvicente e Minardi (1998); Brito, Assaf Neto e Corrar (2009); Carvalho et al (2010). Índice de cobertura de juros. Mede a capacidade de uma empresa gerar lucro operacional suficiente para cobrir suas despesas com juros.

- X_{11} – Lucro Líquido/Receita Líquida [LlRl]

Indicador presente em Elizabetsky (1976); Minussi, Damascena e Ness Júnior (2002); Castro Júnior (2003); Onusic et al (2006); Brito, Assaf Neto (2009); Horta (2010). Indicador de rentabilidade que representa o percentual de lucro da empresa em relação ao seu faturamento. Denominado Retorno sobre Vendas.

- X_{12} – Lajir/Exigível Total [LajirExgt]

Indicador presente em Guimarães e Moreira (2008). Indicador de endividamento que mede a proporção do fluxo de caixa da empresa em relação a suas obrigações. Quanto maior o resultado desse quociente, menor será a probabilidade de uma empresa apresentar dificuldades relacionadas a compromissos financeiros (Guimarães e Moreira, 2008).

- X_{13} – Disponibilidades/Passivo Circulante [DispPc]

Índice presente em Brito, Assaf Neto e Corrar (2009); Aita, Zani e Silva (2010); Horta (2010). Índice de liquidez que representa a capacidade de pagamento que uma empresa

possui levando em consideração apenas suas disponibilidades imediatas e aplicações de curtíssimo prazo. Denominado Índice de Liquidez Imediata.

- X_{14} – Estoques/Ativo Total [EstAt]

Índice encontrado em Elizabetsky (1976); Lanchtermacher e Espenchitt (2001); Minussi, Damascena e Ness Júnior (2002); Brito, Assaf Neto e Corrar (2009). Indicador que mostra o volume de estoques da empresa em relação ao seu ativo total.

- X_{15} – Disponibilidades/Ativo Permanente [DispAp]

Índice visto em Elizabetsky (1976); Lanchtermacher e Espenchitt (2001). Índice que relaciona a parcela dos recursos disponíveis em curtíssimo prazo, com aqueles investidos em bens e direitos de longuíssimo prazo.

- X_{16} – Ativo Permanente/Patrimônio Líquido [ApPl]

Índice encontrado em Brito, Lachtermacher e Espenchitt (2001); Castro Júnior (2003); Brito, Assaf Neto e Corrar (2009). Mede a parcela dos recursos que encontra-se comprometida com o Ativo Permanente.

3.4 Análise dos dados

O primeiro passo da análise dos dados foi obter estatísticas descritivas buscando descrever as empresas, tendo por base os critérios de emparelhamento das empresas, ou seja, setor e tamanho do ativo, utilizando para esse fim, tabelas, histogramas, medidas de tendência central e variabilidade, buscando mostrar a distribuição das empresas dentro de cada um dos grupos. Posteriormente foram construídas algumas medidas estatísticas para descrever o comportamento das variáveis dentro de cada um dos grupos, sendo os resultados obtidos apresentados de forma de tabela.

Os 16 indicadores das 87 empresas foram submetidos então a uma Análise Fatorial, que, segundo Hair *et al* (2005), é um nome genérico dado a uma classe de métodos estatísticos cujo propósito é definir a estrutura subjacente em um conjunto de dados. A Análise Fatorial analisa a estrutura das correlações existentes entre as variáveis e define dimensões latentes comuns, denominadas fatores. Corrar *et al* (2012) afirma que um raciocínio subjacente dessa

técnica implica que se cada variável age de forma independente das demais, existirão tantas dimensões quanto a quantidade de variáveis, no entanto, se houverem relações de dependência entre as variáveis, poderão ser observadas dimensões em quantidade menor, capazes de explicar grande parte da variabilidade dos dados. A aplicação deste método justifica-se pois não é desejável que variáveis com forte correlação com outras variáveis sejam inclusas no modelo. Esse fenômeno é conhecido como multicolinearidade, o qual influência nos erros padrões dos coeficientes, fazendo com que sejam menores, o que dificulta a estimação dos parâmetros do modelo. Como a Análise Fatorial parte do pressuposto que variáveis altamente correlacionadas geram agrupamentos, esse método pode ser empregado para evitar problemas de multicolinearidade.

A análise foi aplicada à base de dados utilizando o comando `principal()`, presente no pacote *psych* do R. O comando permite que seja utilizada a rotação fatorial, que é uma ferramenta bastante importante para a interpretação dos fatores. Segundo Hair *et al* (2005), quando a análise é executada sem rotação, os fatores são extraídos na ordem de importância, assim o primeiro tende a acumular as variáveis com carga significante. Os fatores restantes são calculados com base na quantidade residual de variância, assim, cada fator subsequente tem porções sucessivamente menores de explicações. Ao se aplicar um método de rotação, a variância dos primeiros fatores é redistribuída entre os posteriores, visando atingir um padrão fatorial mais significativo. O método de rotação empregado neste estudo foi o *varimax*, que é um dos mais populares e busca minimizar a quantidade de variáveis em um agrupamento, o que maximiza a variação dos pesos de cada fator, daí seu nome *varimax*.

O número de fatores foi escolhido de acordo com o critério da raiz latente. De acordo com Hair *et al* (2005) esse é o critério mais utilizado para a definição da quantidade de fatores e baseia-se no fato de que cada fator individual deve explicar pelo menos uma variável, assim a análise é feita enquanto os fatores possuem autovalores maiores do que um. A escolha das variáveis foi feita de acordo com o critério da variável substituta, isto é, dentro de cada fator foi escolhida a variável com maior poder de explicação, sendo descartadas as demais.

Para que a Análise Fatorial possa ser aplicada é necessário que a matriz dos dados apresente correlações suficientes que torne uma análise desse tipo justificável. Nesse estudo foram empregados dois critérios para averiguar a aplicabilidade do método. O primeiro deles é o teste de esfericidade de Bartlett, que testa a hipótese nula de que a matriz de correlação da amostra é uma matriz identidade, o que tornaria a Análise Fatorial inadequada. Dessa forma é

desejável obter-se valores pequenos para o valor *p*, menores do que o nível de significância utilizado, para que se possa rejeitar a hipótese nula. Esse teste pode ser aplicado no R por meio do comando `cortest.bartlett()`, do pacote *psych*. O segundo critério aplicado para avaliar a adequação do modelo fatorial à base de dados é o Kaiser-Meyer-Olkin (KMO), que aponta qual a proporção da variância dos dados pode ser considerada comum a todas as variáveis. A análise é aplicável quando o valor observado do KMO é superior à 0,5, sendo esse valor tão melhor quanto mais próximo de 1 ele seja. A medida do KMO foi implementada por meio de uma adaptação para o R do código criado por Trujillo-Ortiz (2006) para o MatLab. A adaptação foi feita por Jay Kerns em 2007.

Depois de definidas as variáveis, o próximo passo foi a aplicação dos modelos de classificação pretendidos, na ordem em que foram apresentados na seção de revisão da literatura: LDA, RL, kNN, CART e ANN.

A normalidade multivariada foi testada no R por meio do comando `mshapiro.test()`, constante no pacote *mvnormtest*, que executa um teste de Shapiro-Wilk, cuja hipótese nula é a de o vetor das variáveis segue uma distribuição normal multivariada, ou seja, é desejável que se obtenha um valor *p* superior ao nível de significância desejado, afim de que não haja evidências para rejeitar a hipótese nula. O nível de significância adotado neste estudo foi de 0,05.

A igualdade das matrizes de variância-covariância foi testada por meio do comando `fligner.test()`, disponível no pacote *stats*. O comando executa o teste de Fligner-Killeen, que testa a hipótese nula de que as variâncias de cada grupo, neste caso, solventes e insolventes, são idênticas. Assim, é desejável obter um valor de p superior ao nível de significância adotado, para que não haja evidências nos dados para rejeitar a hipótese nula.

Depois de testadas as hipóteses de normalidade multivariada e homocedasticidade, prosseguiu-se com a LDA. O processo de estimação da função discriminante começa com a seleção das variáveis que compõe o modelo final (CASTRO, 2003). Uma das maneiras mais conhecidas de encontrar a melhor combinação de variáveis para a otimização do modelo é o procedimento *stepwise*, que realizado no R para análises discriminantes por meio do comando `stepclass()`, do pacote *klaR*. Por fim, utilizou-se o comando `lda()` do pacote *MASS* para criar um modelo baseado em LDA para a previsão de insolvência.

A próxima técnica empregada na classificação das empresas foi a RL. Diferente da LDA, não requer testes preliminares de normalidade e homocedasticidade. Também foi

utilizada um procedimento *stepwise* para a definição de qual combinação de variáveis geraria o melhor modelo. O comando para o *stepwise* adequado ao modelo de Regressão Logística é o `stepwise()`, que pode ser encontrado no pacote *Rcmdr*. O *stepwise* para a RL é feita pelo critério conhecido como *Akaike Information Criterion* (AIC), que baseia-se na função de log-verossimilhança com a introdução de um fator de correção que penaliza conforme a complexidade do modelo (Pedro, 2001). A RL pode ser aplicada adicionando-se o argumento `family=binomial(link=logit)` ao comando `glm()`, o que adapta o modelo linear generalizado (GLM) ao caso particular de em uma regressão logística.

A aplicação do método de kNN no R dá-se por meio do comando `knn()`, disponível no pacote *class*. Para a utilização do método é necessário determinar quantos vizinhos serão considerados para a classificação das observações, e para a estimação do melhor valor, o R dispõe do comando `tune.knn()`, que busca dentro do intervalo definido pelo observador, qual o número de vizinhos ideal. O comando pode ser adicionado ao programa por meio da instalação do pacote *e1071*. Como a identificação desse parâmetro ideal é feita utilizando uma validação cruzada de 10 grupos, os quais são tomados aleatoriamente, a estratégia adotada foi a de rodar o comando `tune.knn()` diversas vezes, tomando como parâmetro para a classificação a moda dos resultados obtidos.

O método das CART foi aplicado ao estudo por meio do comando `rpart()`, o qual pode ser utilizado tanto para classificação como para regressão. Como a intenção deste estudo é classificar as empresas em solventes e insolventes, utilizou-se o comando para construir a árvore de classificação, o qual pode ser feito adicionando o argumento `method="class"`. O pacote necessário para a execução desse programa chama-se *rpart*.

As ANN podem ser implementadas no R com o uso do comando `nnet()`, disponível no pacote *nnet*. Com o argumento `size`, é possível determinar a quantidade de unidades de processamento na camada escondida da rede. O melhor tamanho pode ser obtido com a utilização do comando `tune.nnet()`, que funciona de maneira semelhante ao `tune.knn()` e pode ser instalado com o mesmo pacote (*e1071*). Como o funcionamento dos dois é semelhante, utilizou-se a mesma estratégia definida para o comando anterior a fim de determinar o melhor número de unidades para a camada oculta das ANN. Nesse trabalho as ANN foram geradas com uma adaptação do comando `nnet()`, o `nnetrandom()`, que gera vários modelos gerados pelo `nnet()` e salva o melhor resultado. O comando está disponível no pacote *Biodiversity1*.

Para validar os resultados obtidos, foi implementada uma estratégia de validação cruzada conhecida como *leave one out*. O método consiste em retirar uma observação da amostra, e então utilizar o novo conjunto de dados resultante para a estimação do modelo. A observação que foi retirada é então utilizada para validar o modelo gerado. O processo é repetido até que todas as observações tenham sido utilizadas para a etapa de validação. O resultado mostrado para todos os modelos considera a aplicação da validação cruzada por *leave one out*.

3.5 Comparação dos resultados

Como o estudo propõe-se a avaliar o desempenho dos modelos, uma estratégia de comparação entre os métodos utilizados faz-se necessária. É comum que se utilizem as precisões gerais obtidas para cada modelo construído, como medidas de desempenho dos mesmos. Esse tipo de análise, no entanto, não é indicado para os casos em que existam desproporções entre as classes que compõem a amostra, situação evidenciada nesse trabalho. É recomendada nesta situação a análise da precisão dos modelos em cada grupo (solventes/insolventes).

Para fazer esse tipo de comparação, é possível utilizar curvas ROC, do inglês *Receiver Operating Characteristic*. Esse tipo de análise baseia-se na relação entre a sensibilidade e a especificidade de classificadores binários. A sensibilidade é a taxa de verdadeiros positivos (*true positive rate*), que nesse estudo é a probabilidade de uma empresa solvente ser classificada como tal. O segundo atributo, a especificidade, é a taxa de verdadeiros negativos (*true negative rate*), a probabilidade de uma empresa insolvente ser classificada nesse grupo.

De acordo com Flach (2010), se um modelo de classificação estima um escore que seja proporcional ao grau de certeza com o qual determinada entrada pertença à classe positiva, neste caso, de ser solvente, é possível determinar vários pontos de corte, os quais definirão diferentes proporções para os valores de sensibilidade e especificidade dos modelos. Observando todos os possíveis pontos de corte desde zero até um e ligando-os todos, forma-se uma “curva” composta por segmentos de reta, a qual recebe o nome de Curva ROC.

Como exemplo, pode se considerar o conjunto de dados da Tabela 6. Nela estão dispostos a classe de dez observações e o seu respectivo escore obtido para um modelo de classificação que possua saídas binárias. Se considerarmos como ponto de corte o escore mais elevado, 0,95, todos os valores da tabela serão classificados como negativos, assim o número de falsos positivos será zero, já que nenhum dado negativo foi classificado como positivo, bem como o número de verdadeiros positivos também será zero, já que todas as classes positivas receberam valor negativo para o ponto de corte escolhido.

Tomando agora como ponto de corte o valor seguinte, 0,87, é fácil perceber que o número de falsos positivos continuará zero, enquanto que o número de verdadeiros positivos será um (a observação cujo escore obtido é de 0,95). Assim, a taxa de verdadeiros positivos passará a ser 1/6, aproximadamente 0,17.

Pode usar-se em seguida, como ponto de corte, o valor 0,54. Dessa forma, todos os pontos situados acima deste valor serão classificados como positivos, sendo um deles falso e os demais, verdadeiros. Dessa forma, a taxa de verdadeiros positivos sobe para 3/6, enquanto a taxa de falsos positivos agora passará a ser de 1/4, isto é, 0,5 e 0,25, respectivamente.

Tabela 6 - Exemplo de construção de uma curva ROC

Classe	Escore
Positiva	0,95
Positiva	0,87
Positiva	0,78
Negativa	0,67
Positiva	0,54
Positiva	0,51
Negativa	0,43
Positiva	0,30
Negativa	0,21
Negativa	0,09

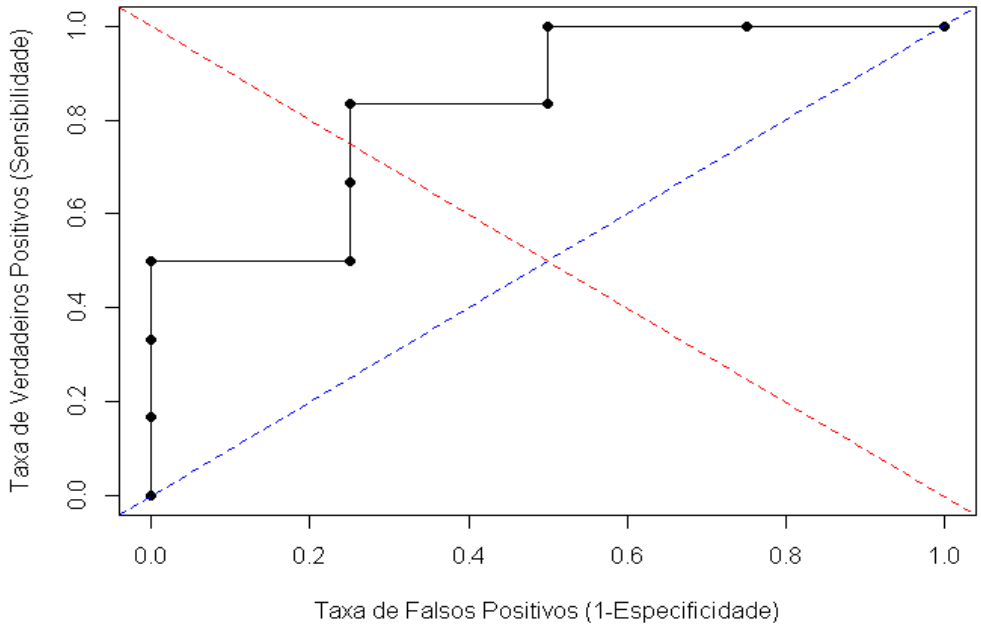
Fonte: Elaborada pelo autor.

Fazendo isso para todos os pontos de corte possíveis, até que todos os dados sejam classificados como positivos, e então construindo um gráfico no qual o eixo das abcissas representa a taxa de falsos positivos, e o das ordenadas contém a taxa de verdadeiros positivos, é possível obter uma curva de segmentos (cor preta) como a do Gráfico 1, a curva ROC. A principal medida de desempenho que se pode obter desse gráfico é o que se chama de AUC, do inglês *Area Under the Curve*, que consiste em calcular o valor da área abaixo da curva. A AUC pode ser interpretada como a probabilidade de uma observação positiva retirada aleatoriamente

receba uma escore maior do que uma observação negativa retirada da mesma maneira. Quanto melhor o ajustamento do modelo ao conjunto de dados, maior será a AUC, que varia entre zero e um. Esta foi a medida adotada neste trabalho para avaliar o desempenho das técnicas utilizadas para a classificação dos dados.

Prati, Batista e Monard (2008) citam outras propriedades interessantes da curva ROC. A linha tracejada em azul representa um modelo de comportamento estocástico, isto é, aleatório. Nele, cada ponto pode ser obtido assumindo que a classe positiva possui probabilidade *p*, enquanto a classe negativa possui probabilidade complementar. Os pontos no quadrante superior esquerdo à essa linha possuem desempenho melhor. Já os pontos que se encontram no quadrante inferior esquerdo em relação à linha tracejada em vermelho, representam modelos que possuem desempenho superior na classe negativa, enquanto os pontos do quadrante superior direito representam modelos cujo desempenho na classe positiva é melhor.

Gráfico 1 - Exemplo de uma curva ROC



Fonte: Elaborada pelo autor.

As curvas ROC podem ser construídas para s cinco modelos aplicados neste trabalho. Para os modelos de LDA, RL e ANN, serão consideradas as probabilidades *a posteriori* de que uma observação pertença ao grupo de empresas solventes, procedimento semelhante ao que é feito para as CART. No caso do método de kNN, a probabilidade de uma observação pertencer a uma determinada classe é dada pela proporção de vizinhos pertencentes a essa classe, em relação à quantidade total de vizinhos considerados para a classificação.

Para a criação das curvas ROC são necessários dois comandos disponíveis no pacote *ROCR*. O primeiro deles é o `prediction()`, que calcula o número de verdadeiros e falsos negativos e positivos, bem como suas taxas. O segundo comando chama-se `performance()`, com o qual podem ser extraídas informações como a razão entre a taxa de verdadeiros e falsos positivos, por meio dos argumentos `measure="tpr"` e `x.measure="fpr"`, para que depois seja construída a curva ROC usando o comando `plot()`. O comando `performance()` permite também que seja aferido o valor da AUC, bastando apenas utilizar o argumento `measure="auc"`.

Além das curvas ROC e da medida AUC, foi construída uma tabela que resume o percentual de acerto de todos os modelos dentro de cada um dos grupos, bem como a precisão para todas as empresas.

4. Resultado

Este capítulo apresenta os resultados obtidos após a adoção dos procedimentos descritos na metodologia. Inicialmente é exibida uma análise das características gerais do banco de dados, em seguida os resultados obtidos para cada modelo são expostos e comentados na mesma ordem em que foram listados no referencial teórico. Por fim, é feita uma comparação do desempenho atingido por cada método, a fim de apontar aquele com melhor performance.

4.1 Análise descritiva dos dados

A análise descritiva é a fase inicial do processo de estudo dos dados coletados. Por meio de estatísticas descritivas é possível organizar, resumir ou descrever aspectos importantes sobre as características de um conjunto de dados. A descrição dos dados também visa identificar anomalias, que podem ser resultante de registros incorretos, e também dados dispersos que fujam à tendência do restante do conjunto (Reis e Reis, 2002).

As primeiras análises feitas buscaram extrair mais informações sobre a distribuição das empresas tendo como base a classificação setorial e o tamanho do ativo, que foram os critérios adotados para a seleção da amostra. A Tabela 7 traz as informações quanto aos setores, enquanto a Tabela 8 e a Figura 4 demonstram a análise de acordo com o tamanho do ativo.

Tabela 7 - Distribuição das empresas por setor

Subsetores	Insolventes		Solventes		Total	
Agroindústria	1	4,76%	4	6,06%	5	5,75%
Comércio de máquinas e veículos pesados	1	4,76%	2	3,03%	3	3,45%
Construção e engenharia	2	9,52%	7	10,61%	9	10,34%
Eletroeletrônicos	1	4,76%	5	7,58%	6	6,90%
Energia elétrica	2	9,52%	10	15,15%	12	13,79%
Madeira e papel	1	4,76%	4	6,06%	5	5,75%
Material de construção	1	4,76%	1	1,52%	2	2,30%
Material de transporte	1	4,76%	5	7,58%	6	6,90%
Metalurgia	1	4,76%	4	6,06%	5	5,75%
Química	1	4,76%	4	6,06%	5	5,75%
Tecidos, vestuário e calçados	8	38,10%	18	27,27%	26	29,89%
Transportes aéreos	1	4,76%	2	3,03%	3	3,45%

Total	21	100,00%	66	100,00%	87	100,00%

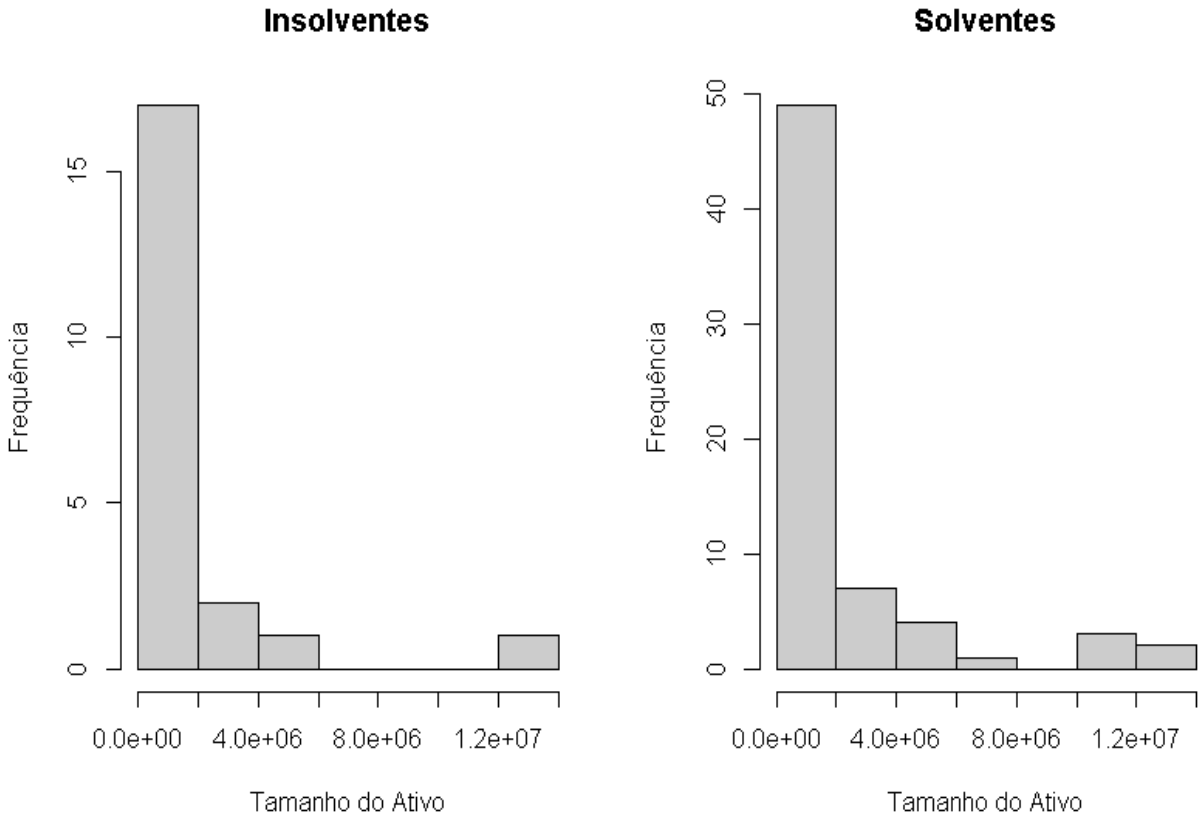
Fonte: Elaborada pelo autor.

Tabela 8 – Medidas do tamanho do Ativo Total por grupo

	Insolventes	Solventes	Geral
Mínimo	23.620	68.216	23.620
Máximo	12.935.830	13.662.280	13.662.280
Média	1.227.364	2.053.753	1.854.280
Mediana	182.075	574.097	506.987
Desvio	2.915.823	3.207.672	3.143.312

Fonte: Elaborada pelo autor.

Figura 4 - Histograma da distribuição do Ativo Total por grupos



Fonte: Elaborada pelo autor.

É possível perceber que, levando-se em conta os critérios adotados para a composição da amostra, apesar da divergência no tamanho dos grupos, as empresas distribuem-se de forma mais ou menos homogênea dentro do grupo ao qual pertencem.

Continuando a análise descritiva, o procedimento seguinte envolveu o cálculo de algumas medidas descritivas das variáveis independentes. Foram calculadas medidas de tendência central, variabilidade e de posição para os 16 indicadores contábeis. As Tabelas 9 e

10 apresentam essas medidas para os grupos de empresas solventes e insolventes, respectivamente.

Tabela 9 - Estatísticas descritivas das variáveis (empresas insolventes)

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Amplitude Total	Desvio Padrão
X1	-8,6140	-0,9828	-0,3911	-1,0820	-0,1699	0,2511	8,8651	1,9655
X2	-0,4126	-0,2116	-0,0855	-0,1015	0,0703	0,2221	0,6347	0,1959
X3	-0,8991	-0,6314	-0,3199	-0,1847	0,3032	0,6141	1,5132	0,5162
X4	0,0372	0,4648	0,5885	0,7000	0,7439	2,7030	2,6658	0,5410
X5	0,6195	0,7674	1,4700	2,2160	2,7120	9,9060	9,2865	2,2178
X6	0,0422	0,1597	0,2883	0,4680	0,5604	2,0080	1,9658	0,4719
X7	-1,2600	-0,5061	-0,1941	-0,3447	-0,0532	0,0374	1,2974	0,3692
X8	0,0218	0,1136	0,1797	0,3595	0,4860	1,8180	1,7962	0,4235
X9	0,0246	0,1520	0,3102	0,4534	0,6084	1,4310	1,4064	0,4057
X10	-6,9960	-0,9498	-0,4148	-0,6501	0,1819	1,3840	8,3800	1,6619
X11	-13,6100	-0,7351	-0,4766	-1,3460	-0,0884	0,0677	13,6777	3,1303
X12	-0,6144	-0,1067	-0,0422	-0,0548	0,0395	0,1999	0,8143	0,1683
X13	0,0001	0,0012	0,0053	0,0292	0,0410	0,1437	0,1436	0,0424
X14	0,0028	0,0151	0,0593	0,0938	0,1309	0,4169	0,4141	0,1086
X15	0,0001	0,0038	0,0115	0,0586	0,0937	0,4166	0,4165	0,1009
X16	-1,8600	-0,3432	-0,1067	0,8991	1,6820	9,3390	11,1990	2,5296

Fonte: Elaborada pelo autor.

Tabela 10 - Estatísticas descritivas das variáveis (empresas solventes)

Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Amplitude Total	Desvio Padrão
X1	-1,2050	0,0379	0,2254	0,1603	0,3166	0,5178	1,7228	0,2466
X2	-0,2735	0,0365	0,0819	0,0943	0,1237	0,5285	0,8020	0,1076
X3	-0,5111	0,5540	0,8002	1,0400	1,3160	4,8130	5,3241	0,8883
X4	0,0120	0,5192	0,7774	0,8676	1,1400	2,1310	2,1190	0,4624
X5	0,1719	0,4216	0,5350	0,5633	0,6186	2,0450	1,8731	0,2713
X6	0,2138	1,1340	1,7180	1,7830	2,1770	4,8790	4,6652	0,9159
X7	-0,7526	0,0095	0,0421	0,0409	0,0761	0,4263	1,1789	0,1279
X8	0,1626	0,9797	1,1900	1,3210	1,5430	4,0830	3,9204	0,7213
X9	0,2348	0,7462	1,1780	1,2080	1,4570	3,5490	3,3142	0,6407
X10	-2,1990	0,6054	1,4660	2,0720	3,1560	8,3080	10,5070	2,1938
X11	-0,9729	0,0248	0,0621	0,0468	0,0998	0,6375	1,6104	0,1762
X12	-0,1552	0,0601	0,1742	0,1893	0,2822	0,8361	0,9913	0,1974
X13	0,0003	0,0564	0,2090	0,3664	0,3865	2,7940	2,7937	0,5365
X14	0,0007	0,0373	0,1270	0,1227	0,1818	0,4303	0,4296	0,0975
X15	0,0003	0,0355	0,1401	0,4736	0,4016	9,5020	9,5017	1,3907
X16	-3,3020	0,5701	0,7953	1,0500	1,3490	6,2930	9,5950	1,2892

Fonte: Elaborada pelo autor.

Ao analisar os valores mínimos, máximos e a amplitude total das variáveis, em comparação com suas respectivas médias, medianas e quartis, é fácil perceber que existem

outliers, especialmente no grupo de empresas insolventes. No entanto, optou-se pela manutenção tanto das variáveis, quanto das empresas que apresentaram esse tipo de observação, com o objetivo de testar a robustez dos modelos a esses dados.

4.2 Análise fatorial

Após a investigação preliminar dos dados, deu-se então início à AF. Inicialmente foi testada a adequação do conjunto ao modelo fatorial por meio do teste de esfericidade de Bartlett e do KMO. O resultado de ambos pode ser visto na Tabela 11.

Tabela 11 - Resultado dos testes de Bartlett e KMO

Teste	Bartlett		KMO
	χ^2	p-valor	
Resultado	1.409,1560	0,0000	0,6201

Fonte: Elaborada pelo autor.

Os valores obtidos indicam que é possível aplicar a AF ao conjunto de dados, apresentando um valor *p* inferior a 0,0001 e um KMO de 0,6201, considerado mediano segundo os padrões definidos por Kaiser (1974). Dessa forma, prosseguiu-se com a análise. O critério da raiz latente levou à escolha de oito fatores, dos quais retirou-se a variável com maior escore. A Tabela 12 mostra a composição de cada fator, bem como a carga fatorial (após rotação *varimax*) cada variável dentro do fator que ela compõe, postas em ordem decrescente.

Tabela 12 - Resultado da Análise Fatorial

Fator	Variável	Código	Carga fatorial
1	**X_{12}**	**LajirExgt**	**0,89575**
	X_{10}	LajirDespFin	0,86318
	X_2	LajirAt	0,84238
	X_4	RlAt	0,63512
2	**X_1**	**AcPcAt**	**0,93895**
	X_5	ExgtAt	-0,85529
	X_7	LlAt	0,73641
3	**X_{13}**	**DispPc**	**0,95126**
	X_8	AcEstPc	0,78774
	X_6	AcPc	0,7753

4	**X_3**	**PlExgt**	**0,77756**
	X_9	AcRlpExgt	0,7738
5	**X_{11}**	**LlRl**	**0,90331**
6	**X_{14}**	**EstAt**	**0,9703**
7	**X_{15}**	**DispAp**	**0,97757**
8	**X_{16}**	**ApPl**	**0,95502**

Fonte: Elaborada pelo autor.

O total da variância explicada pelos oito fatores é de 92,10%. Todos os valores obtidos para as comunalidades foram superiores a 0,5 (o menor valor observado foi de 0,7492 para a variável X_4).

Nota-se que as variáveis foram agrupadas segundo as grandezas utilizadas na composição do indicador. O primeiro fator agrupou todas as variáveis com LAJIR, e também a relação entre Receita Líquida e Ativo Total, mas com uma carga fatorial consideravelmente menor. O segundo fator agrupa os indicadores com o Ativo Total. Já as variáveis com Passivo Circulante em sua composição estão agrupadas no terceiro fator. Duas variáveis com Exigível Total estão no fator 4. As demais variáveis ficaram em fatores separados.

Assim, apenas as variáveis X_1, X_3, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} e X_{16} passaram a ser consideradas para a aplicação dos modelos de previsão de insolvência.

4.3 Modelos de previsão de insolvência

Nesta seção apresentam-se os resultados obtidos para cada um dos modelos aplicados no trabalho, seguindo a mesma ordem do referencial teórico.

4.3.1 Análise discriminante linear

Os resultados dos testes de normalidade e de homocedasticidade apontaram para a não adequação da LDA aos dados utilizados. O teste de normalidade de Shapiro-Wilk resultou em p-valores da ordem de 10^{-7} e 10^{-13} para o grupo de empresas insolventes e solventes, respectivamente, fazendo com que a hipótese nula de normalidade fosse rejeitada. A Tabela 13

traz os resultados do deste de Shapiro-Wilk para a normalidade multivariada, e univariada de todos os indicadores.

Tabela 13 - Resultados do teste de Shapiro-Wilk

Variáveis	Shapiro-Wilk			
	Insolventes		Solventes	
	W	*p*-valor	W	*p*-valor
Multivariada (X_1, X_3, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16})	0,50760	0,00000	0,50620	0,00000
X_1	0,58970	0,00000	0,79550	0,00000
X_3	0,92780	0,11020	0,86850	0,00001
X_{11}	0,43600	0,00000	0,68640	0,00000
X_{12}	0,86720	0,00693	0,91480	0,00027
X_{13}	0,39490	0,00000	0,63480	0,00000
X_{14}	0,81790	0,00097	0,93340	0,00172
X_{15}	0,64160	0,00000	0,31120	0,00000
X_{16}	0,76220	0,00014	0,83570	0,00000

Fonte: Elaborada pelo autor.

Tentou-se então utilizar transformações sugeridas por Hair *et al* (2005) que têm por objetivo normalizar variáveis não-normais, como por exemplo, potenciação, radiciação e aplicação de logaritmo. Para isso, foi observado inicialmente quais das variáveis eram normais quando analisadas isoladamente dentro de cada grupo, e depois foram aplicadas transformações monótonas a cada uma das variáveis não-normais. Apenas algumas variáveis obtiveram resultado positivo para o teste de Shapiro-Wilk após as transformações. Mesmo assim, considerando apenas as variáveis normais no teste de normalidade multivariada, a hipótese continuou sendo rejeitada.

A homocedasticidade multivariada também não foi observada, segundo os resultados do teste de Flinger-Killeen. Os resultados obtidos para o resultado considerando todas as variáveis, bem como cada uma delas individualmente encontra-se na Tabela 14. Hair *et al* (2005), sugere que as mesmas transformações utilizadas para atingir a normalidade também podem servir para a homocedasticidade. Essas transformações, no entanto, não conseguiram levar o conjunto de dados à observância desse pressuposto.

Tabela 14 – Resultados do teste de Fligner-Killeen

Variáveis	Fligner-Killeen p-valor
Multivariada (X1, X3, X11, X12, X13, X14, X15, X16)	0,00000
X_1	0,00009
X_3	0,47680
X_{11}	0,00000
X_{12}	0,29430
X_{13}	0,00000
X_{14}	0,30640
X_{15}	0,00027
X_{16}	0,11950

Fonte: Elaborada pelo autor.

Apesar dos resultados dos testes preliminares indicaram que a LDA não é uma técnica adequada aos dados, ainda assim insistiu-se na utilização do método, devido à sua grande ocorrência na literatura acerca da previsão de insolvência, avaliando-se assim o desempenho deste método, mesmo em situação de violação dos pressupostos.

Por meio do método *stewipse* chegou-se à conclusão de que as variáveis X_3 e X_{12}, formam o melhor modelo para discriminar os dois grupos em estudo. Vale ressaltar que as duas variáveis atenderam ao critério de homocedasticidade e obtiveram os maiores resultados para o teste de normalidade. A Tabela 15 mostra um quadro com o resumo da classificação gerada pelo método. A classificação de todas as empresas para este e todos os outros métodos encontra-se no Apêndice B deste trabalho. A função discriminante do modelo é:

$$D = 0{,}8836478X_3 + 2{,}9878692X_{12}$$

Tabela 15 – Tabela de classificação do modelo LDA

	Classificadas como:		Total
	Insolvente	Solvente	
Insolvente	15	6	21
	71,43%	28,57%	100,00%
Solventes	1	65	66
	1,52%	98,48%	100,00%
Total	16	71	87

Fonte: Elaborada pelo autor.

O modelo da LDA acertou a classificação de 80 empresas, o que representa uma precisão de 91,95%. No grupo das empresas insolventes o modelo foi capaz de classificar corretamente 15 das 21 empresas, indicando 71,43% de precisão. Já para as empresas solventes,

o modelo classifica corretamente 65 das 66 companhias que compõem este grupo, isto é, 98,48%.

4.3.2 Regressão logística

O segundo modelo aplicado no estudo foi o de RL. O processo de *stepwise* para o método indicou que as variáveis X_3 e X_{12}, assim como na LDA, são as mais úteis para a classificação do conjunto de dados. A Tabela 16 apresenta os coeficientes estimados pelo modelo, bem como seus valores exponenciais, as estatísticas de Wald e os valores *p*.

Tabela 16 - Resultados da Regressão Logística

Variáveis Independentes	Coeficiente (B)	Exp(B)	Wald	valor *p*
X_3	2,8483	17,2584	3,31	0,0009
X_{12}	7,1560	1.281,7736	2,17	0,0301
Constante	-0,3193	0,7267	-0,70	0,4865

Fonte: Elaborada pelo autor.

Observando os valores gerados pelo modelo, é possível determinar que o ponto de corte que maximiza o acerto do modelo situa-se próximo a 0,65. O Gráfico 2 demonstra a situação graficamente. Uma síntese dos resultados obtidos pode ser vista na Tabela 17.

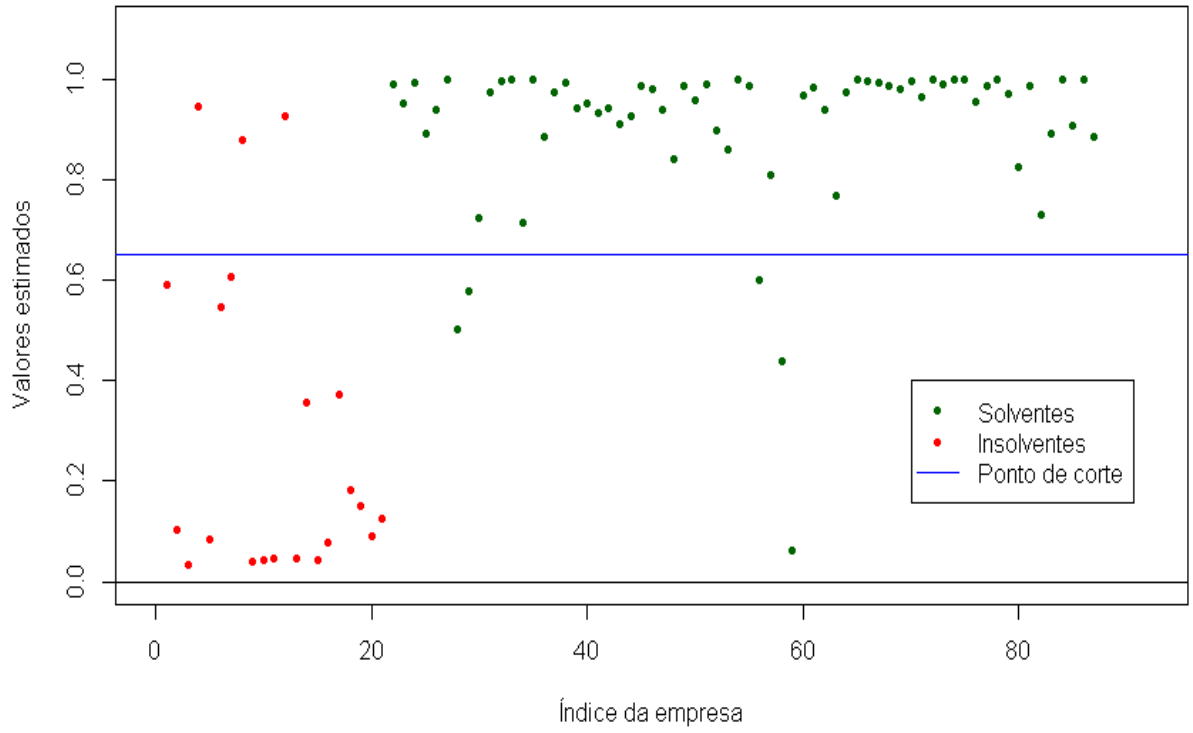
Tabela 17 - Tabela de classificação do modelo RL

	Classificadas como:		Total
	Insolvente	Solvente	
Insolvente	18	3	21
	85,71%	14,29%	100,00%
Solventes	5	61	66
	7,58%	92,42%	100,00%
Total	23	64	87

Fonte: Elaborada pelo autor.

O modelo de RL acertou a classificação de 79 das 87 empresas, o que representa 90,80% do conjunto de dados. No grupo das empresas insolventes o modelo classificou de maneira correta 18 das 21 empresas, indicando um percentual de precisão de 85,71%. Já no grupo de empresas solventes, o método acertou a predição de 61 das 66 empresas, o que lhe confere percentual de acerto de 92,42%.

Gráfico 2 - Valores estimados e ponto de corte



Fonte: Elaboração do autor.

O valor do VIF para os coeficientes estimados foi de 1,0569 para ambos (o que representa um valor de R^2 de 0,0538 para um modelo de regressão em que uma variável independente fosse usada para explicar outra), indicando baixa correlação entre as variáveis utilizadas para a RL e um bom ajustamento do modelo. Com o comando `summary()` é possível ver se as variáveis possuem significância estatística por meio do valor Z, comumente chamado de estatística de Wald. Os valores *p* obtidos pelo teste de Wald para os indicadores X_3 e X_{12} foram respectivamente 0,000935 e 0,03015, o que mostra que são significantes para um nível de significância de 5%. O coeficiente R^2 de Nagelkere obtido foi de 0,685, o que indica uma boa qualidade do ajustamento do modelo.

4.3.3 Classificador dos vizinhos mais próximos

Na aplicação do kNN, buscou-se inicialmente determinar a quantidade de vizinhos a ser considerada para o modelo. Devido à componente aleatória associada ao comando `tune.knn()`, que é utilizado para resolver o problema do número de vizinhos, o mesmo foi

executado 1000 vezes, considerando um intervalo entre um e dez. O melhor parâmetro obtido segundo esse método indica que um vizinho é a melhor opção, aparecendo 620 vezes no total. O valor que obteve, em seguida, a maior frequência foi o quatro, sendo sugerido pelo comando 113 vezes. A Tabela 18 apresenta a tabela de classificação utilizando um vizinho.

Tabela 18 – Tabela de classificação do modelo kNN

	Classificadas como:		Total
	Insolvente	Solvente	
Insolvente	15	6	21
	71,43%	28,57%	100,00%
Solventes	5	61	66
	7,58%	92,42%	100,00%
Total	20	67	87

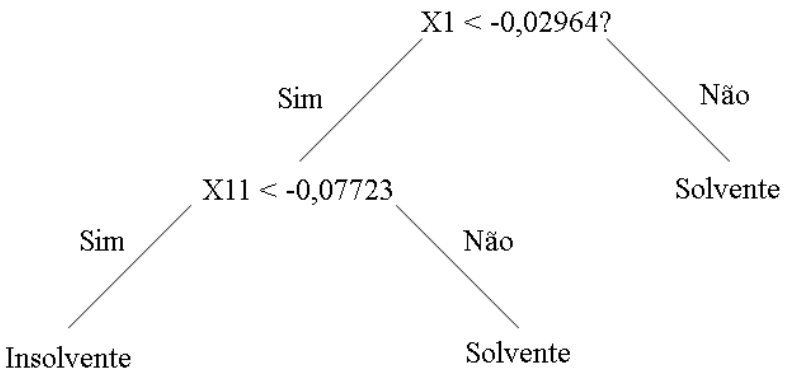
Fonte: Elaborada pelo autor.

O método acertou a classificação de 76 empresas, o que representa 87,36% do total da amostra. No grupo de empresas insolventes, classificou corretamente 15 das 21 companhias, atingindo uma precisão de 71,43%. Já no grupo de empresas solventes, 61 das 66 empresas foram corretamente preditas, o que indica uma acurácia de 92,42%.

4.3.4 Árvores de classificação

O modelo das CART utilizou variáveis diferentes daquelas que foram usadas na LDA e RL. Enquanto os dois últimos modelos basearam sua decisão em índices de endividamento e estrutura (X_3 e X_{12}), as CART utilizaram um índice que retrata a liquidez, a relação entre o Capital Circulante Líquido com o Ativo Total (X_1), e um de rentabilidade, a Margem Líquida (X_{11}). A Figura 5 mostra as regras de classificação adotadas na formação da árvore. A tabela de classificação obtida pelas CART pode ser vista na Tabela 19.

Figura 5 - Árvore de Classificação do modelo CART



Fonte: Elaborada pelo autor.

Tabela 19 – Tabela de classificação do modelo CART

	Classificadas como:		Total
	Insolvente	Solvente	
Insolvente	17	4	21
	80,95%	19,05%	100,00%
Solventes	1	65	66
	1,52%	98,48%	100,00%
Total	16	71	87

Fonte: Elaborada pelo autor.

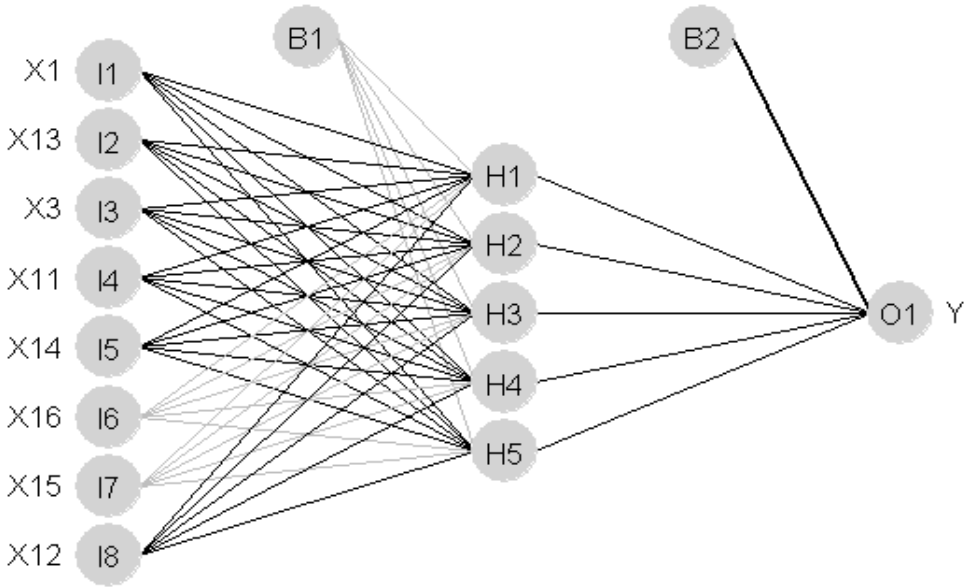
O modelo das CART acertou a classificação de 82 das 87 empresas, indicando uma precisão geral de 94,25%. No grupo de empresas insolventes, o modelo classificou de maneira correta 17 das 21 observações, o que representa 80,95% de acurácia. Já para o grupo de empresas solventes, o método classificou corretamente 65 das 66 observações, o que indica uma precisão de 98,48%.

Aplicando o comando `summary()` à predição feita, é possível ver a importância que cada uma das variáveis tem para a classificação utilizando o algoritmo das CART, mesmo que nem todas elas tenham sido utilizadas para a classificação. O resultado obtido para essa consulta mostra que a variável X_1 tem importância de 27%, X_{11} de 20%, X_3 de 17%, X_{16} possui 12%, X_{13} e X_{15} importam 10% e X_{12} apenas 3%. O indicador X_{14} não foi considerado importante pelo modelo das CART.

4.3.5 Redes neurais

A determinação da quantidade de unidades na camada escondida gerou números que não apontavam uma única quantidade mais clara, as quantidades mais recorrentes situaram-se entre três e nove, sendo cinco a mais observada em 1.000 testes realizados. Resolveu-se então testar todas as quantidades que foram sugeridas e foi observado que os escores obtidos eram semelhantes. Tendo em vista essa situação, a quantidade considerada para construir o modelo de ANN para o estudo foi a de cinco, já que foi a mais comum, mesmo que a diferença dessa observação para as demais não permita afirmar com certeza que este é o valor ideal. A Figura 6 mostra o desenho do modelo de ANN gerado pelo R considerando as oito variáveis do estudo, em que I_n, n =1, 2, 3 ..., 8, são as unidades da camada de entrada, que recebe o valor das variáveis independentes; H_n, n = 1, 2, ..., 5 são as unidades da camada oculta, que recebem os *outputs* da camada de entrada; O_1 é a unidade da camada de saída, que gera o resultado utilizado para a classificação a partir dos *outputs* da camada oculta; Os neurônios B_1 e B_2, são denominados de *bias*, cuja função é aumentar os graus de liberdade, permitindo uma melhor adaptação, por parte da rede neural, ao conhecimento fornecido a ela (SOARES e SILVA, 2011). A Tabela 20 traz um resumo da classificação obtida utilizando as ANN.

Figura 6 - Rede neural gerada para o conjunto de dados



Fonte: Elaborada pelo autor.

Tabela 20 – Tabela de classificação do modelo de ANN

	Classificadas como:		Total
	Insolvente	Solvente	
Insolvente	21	0	21
	100,00%	0,00%	100,00%
Solventes	1	65	66
	1,52%	98,48%	100,00%
Total	16	71	87

Fonte: Elaborada pelo autor.

O modelo das ANN acertou a classificação de 86 das 87 empresas, o que representa 98,85% do conjunto de dados. No grupo de empresas insolventes o modelo não apresentou falhas, classificando-o 100% corretamente. Já no grupo de empresas solventes, o modelo acertou em 65 das 66 observações, lhe conferindo uma precisão de 98,48% para esse grupo.

4.4 Comparação dos modelos de previsão

Com todos os modelos construídos e dispostos todos os resultados, antes que se procedesse com a comparação dos modelos, verificou-se quais empresas haviam sido classificadas de maneira equivocada mais vezes e depois buscou-se alguma possível explicação para o problema.

Todos os modelos erraram na classificação da empresa 59, Wiest, pertencente ao grupo de empresas solventes, muito embora alguns deles focassem em indicadores diversos. Ao serem analisadas as Notas Explicativas emitidas pela empresa referentes ao ano de 2005, foi constatado que, apesar de não ter feito pedido de concordata ou recuperação judicial, a empresa reconhece no relatório que naquele ano enfrentou prejuízos contínuos, ocasionados por endividamento financeiro e fiscal, deficiência de capital de giro e baixo índice de liquidez.

A empresa 4, Tecnosolo, empresa pertencente ao grupo de empresas insolventes, foi classificada de forma incorreta em quatro dos cinco modelos, sendo avaliada corretamente apenas pelas ANN, que mesmo assim atribuíram à empresa um escore relativamente alto quando comparado com as demais companhias insolventes. Nas Notas Explicativas de 2011 não havia nenhum indício de que a empresa passava por dificuldades financeiras de qualquer ordem. A empresa inclusive encerrou o ano de 2011 com lucro líquido de R$ 6,9 milhões. As

Demonstrações Financeiras de 2012 ainda não se encontravam disponibilizadas até ao encerramento desse trabalho, por isso não foi possível avaliar as Notas Explicativas de 2012.

A empresa 12, Sansuy, pertencente ao grupo de empresas insolventes, assim como a Tecnosolo, foi classificada incorretamente em quatro dos cinco modelos, e obteve um escore discrepante dentro do grupo das empresas insolventes. No entanto, a análise das Notas Explicativas dos anos de 2004 e 2005 não deu nenhum indício que pudesse esclarecer tal situação.

Depois de feita a investigação dos resultados mais estranhos, deu-se prosseguimento com a análise comparativa dos modelos. O primeiro passo tomado para avaliar o desempenho dos métodos foi comprar os seus percentuais de acerto, ou seja, a precisão para cada grupo. A Tabela 21 traz todas essas informações. Convencionalmente chama-se Precisão Tipo I, aquela associada à classificação de empresas insolventes, enquanto Precisão Tipo II representa a classificação correta das empresas solventes.

Tabela 21 - Precisões obtidas pelos modelos

Modelo	Precisão Tipo I	Precisão Tipo II	Precisão Geral
LDA	71,43%	98,48%	91,95%
RL	85,71%	92,42%	90,80%
kNN	71,43%	92,42%	87,36%
CART	80,95%	98,48%	94,25%
ANN	100,00%	98,48%	98,85%

Fonte: Elaborada pelo autor

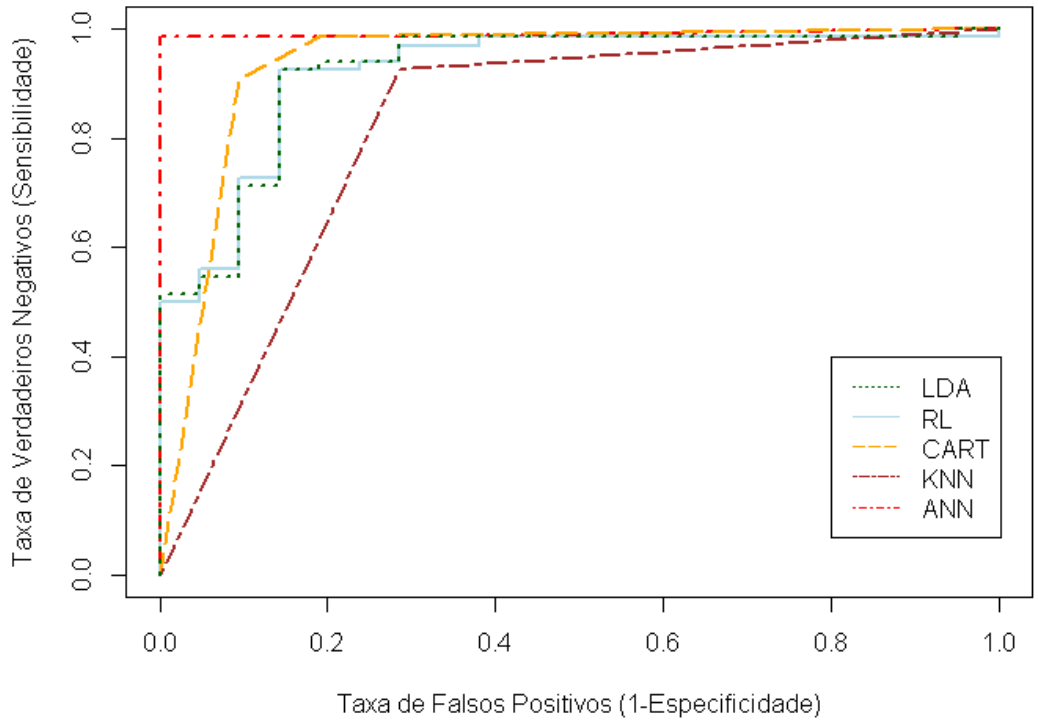
Com base nessa análise, é possível notar que o modelo de ANN obteve o melhor desempenho nos três tipos de precisão, o que corrobora com os resultados obtidos por estudos realizados anteriormente, que indicam que as ANN tem desempenho mais robusto, especialmente quando a amostra utilizada é pequena, como é o caso desta pesquisa. O método errou apenas a classificação de uma empresa no grupo das solventes, exatamente a empresa 59, que, como visto anteriormente, passou por graves dificuldades financeiras no período analisado. O modelo que obteve o segundo melhor desempenho pelo critério adotado foi o das CART, que errou no grupo de empresas solventes apenas a classificação da empresa 59, porém classificou erradamente quatro empresas no grupo das insolventes.

O modelo com pior desempenho foi, sem dúvida, o kNN, que, além de obter a precisão geral mais baixa, também obteve a pior precisão dentro de cada grupo. Não fica muito claro qual o modelo obteve melhor desempenho entre LDA e RL. O primeiro modelo, apesar

de ter obtido um erro geral menor, especialmente se for levado em consideração que o único erro apresentado para o grupo de empresas solventes foi de uma empresa com problemas financeiros, sua precisão pode ter sido afetada pelo desbalanceamento dos grupos, uma vez que seus erros concentram-se nas empresas insolventes, que compõem a menor parte da amostra. Neste sentido, a RL apresenta uma distribuição de erro por grupos menos desproporcional.

Após a análise da precisão, procedeu-se então com a criação das curvas ROC. Para isso, foi feita inicialmente a coleta das probabilidades *a posteriori*, ou seja, das probabilidades preditas de que cada observação pertencesse ao grupo de empresas solventes (classe positiva). Ressalta-se que, como o modelo de kNN considerou apenas um vizinho, as probabilidades dadas pelo método de uma observação ser solvente só poderiam ser 0 ou 1, isto é, o vetor de probabilidades é igual ao vetor de classificação. O Gráfico 3 mostra as curvas ROC obtidas para todos os cinco modelos. O Apêndice D mostra as probabilidades utilizadas na construção das curvas.

Gráfico 3 - Curvas ROC para os cinco modelos



Fonte: Elaborada pelo autor.

Depois de obtidas as curvas ROC, o próximo passo foi a obtenção da AUC. O melhor modelo nesse quesito foi novamente o construído com as ANN, cuja área abaixo da curva (AUC) foi de 0,9906. As CART novamente obtiveram o segundo melhor desempenho, com AUC de 0,9369. O desempenho da LDA e da RL foi bastante parecido, com uma ligeira

vantagem para o primeiro modelo, cuja área calculada foi de 0,9221, enquanto o segundo obteve 0,9199. O único modelo com AUC inferior à 0,9 foi o kNN, cuja área calculada foi de 0,8193. Segundo os critérios normalmente adotados para a análise da AUC, valores acima de 0,9 são tidos como excelentes, enquanto valores entre 0,8 e 0,9 são considerados bons.

5. Conclusão

O objetivo desta pesquisa foi aplicar métodos de classificação para a construção de modelos de predição de insolvência, utilizando índices contábeis de empresas brasileiras de capital aberto. As técnicas empregadas foram Análise Discriminante Linear, Regressão Logística, Classificador dos Vizinhos Mais Próximos, Árvores de Classificação e Redes Neurais Artificiais.

Foram obtidas informações de demonstrações contábeis de 87 empresas, sendo 21 insolventes e 66 solventes, com as quais construíram-se 16 índices que retratavam a liquidez, rentabilidade e endividamento ou estrutura do capital das empresas. Esses indicadores foram submetidos, inicialmente, a uma Análise Fatorial, para que, por meio do critério da variável substitua, ou seja, aquela com maior carga dentro de um fator, o número total de indicadores fosse reduzido, buscando principalmente, evitar problemas de multicolinearidade. O número de variáveis então reduziu para oito.

Prosseguiu-se com a aplicação dos modelos, primeiramente com a Análise Discriminante Linear. Para a utilização desse modelo, os seus pressupostos de normalidade e igualdade das matrizes de variância para os dois grupos foram testados. Embora os resultados desses testes indicassem a inadequação do modelo aos dados, a LDA foi utilizada. Por meio de um procedimento *stepwise*, verificou-se que o melhor modelo discriminante era composto pelas variáveis X_3 e X_{12}, Endividamento Geral e Lajir sobre Capitais de Terceiros. A precisão geral do modelo atingiu 91,95%, sendo 71,43% no grupo de empresas insolventes e 98,48% nas empresas solventes.

O segundo método aplicado foi a Regressão Logística. Adotou-se um procedimento semelhante ao da LDA para que se chegasse no modelo com melhor capacidade preditiva. As variáveis que foram consideradas importantes para a RL foram as mesmas da LDA. O seu desempenho no entanto ficou um pouco abaixo, com 90,80% de precisão geral, 85,71% no grupo insolvente e 92,42% no grupo solvente. Apesar do desempenho geral menor, observou-se que o erro dentro dos grupos era distribuído de maneira mais homogênea.

A terceira técnica aplicada foi o Classificador de Vizinhos Mais Próximos. O primeiro passo adotado para a utilização do método foi a determinação do número ideal de vizinhos, determinado como sendo um. Com esse valor efetuou-se a análise e foi observado

uma precisão geral de 87,36%, a pior observada dentre todos os modelos. A precisão dentro do grupo de insolventes foi de 71,43% e 92,42% para as empresas solventes.

As Árvores de Classificação foram o quarto método aplicado para a construção de um modelo de previsão de insolvência. A técnica utilizou duas divisões em duas variáveis diferentes para chegar a uma classificação com precisão geral de 94,25%, o segundo melhor desempenho considerando esse parâmetro. O modelo acertou a classificação de 80,95% das empresas insolventes e 98,48% daquelas com situação oposta. Os indicadores mais importantes para o modelo foram o Capital Circulante Líquido sobre Ativo Total e a Margem Líquida.

O último método aplicado foram as Redes Neurais Artificiais. O modelo construído foi de múltiplas camadas, com uma camada escondida com cinco unidades de processamento. O valor de cinco unidades foi escolhido de maneira quase arbitrária, já que não ficou evidente qual o número ideal de unidades para a camada escondida. A precisão geral obtida pelas ANN foi a melhor, com 98,85%. O modelo acertou a classificação de todas as empresas do grupo das insolventes, e errou apenas em um caso das solventes, atingindo 98,48% nesse grupo.

Como forma de inferir melhor sobre qual modelo apresentava melhores resultados para o conjunto de dados, foi utilizada uma análise das curvas ROC obtidas pelos métodos. A análise foi efetuada considerando o valor da área abaixo da curva (AUC). Os valores obtidos corroboraram com a análise feita com base nas precisões. O modelo com valor mais elevado para a AUC foram as ANN, com 0,9906. Em seguida vieram as CART com 0,9369. A LDA e a RL obtiveram 0,9221 e 0,9199, respectivamente, desempenho muito semelhante. Todos esses modelos foram classificados como excelentes, segundo o critério da AUC. O único modelo que ficou abaixo desse patamar foi o kNN, com área abaixo da curva de 0,8193, o que ainda assim é considerado um bom valor.

Todas as análises foram feitas por meio da ferramenta estatística R. Os resultados obtidos foram validados por meio do método de *leave one out*, que consiste em fazer uma validação cruzada com todas as observações que compõem a amostra.

Com base nos resultados é possível concluir que modelos de classificação constituem uma ferramenta poderosa para prever problemas financeiros, auxiliando gestores e investidores na tomada de decisão e contribuindo para a redução do risco de crédito.

O modelo baseado em Redes Neurais obteve um desempenho consideravelmente superior aos demais, o que confirma a hipótese sugerida pela literatura de que esse método tem um desempenho superior em amostras pequenas.

Os índices que foram mais importantes, considerando observações feitas a partir da RL e das CART, foram o Capital Circulante Líquido sobre o Ativo Total, Endividamento Geral e Margem Líquida. A primeira destas variáveis é citada também em muitos outros estudos como relevante para a previsão de insolvência (BEAVER, 1966; ALTMAN, 1968; SANVICENTE e MINARDI, 1998), porém cabe ressaltar que não há unanimidade sobre quais indicadores são os mais importantes. Apenas aponta-se que índices de liquidez, auxiliados por índices de endividamento e rentabilidade, são a forma mais apropriada de se construir este tipo de modelo.

Este trabalho apresentou como limitação, o tamanho reduzido da amostra, que contava com apenas 87 empresas. Outro ponto a ser destacado foi a ausência de algumas informações nas demonstrações contábeis o que dificultou ou impossibilitou a construção de alguns indicadores. Outra dificuldade encontrada foram as mudanças ocorridas nas normas contábeis ao longo do período estudado, que implicaram, não raro, em mudanças na estrutura das demonstrações contábeis.

Existem ainda limitações ligadas à utilização de indicadores contábeis para esse tipo de estudo. A primeira delas é a possibilidade de manipulação por parte dos gestores, o que pode gerar índices que não representam de forma fidedigna a realidade da empresa. Há também o fato de que as demonstrações contábeis de um determinado ano apenas serão divulgadas durante o ano subsequente. Como neste trabalho foram utilizadas demonstrações de um ano anterior à entrada em insolvência, algumas das empresas já haviam feito seu pedido de recuperação judicial antes de divulgarem as informações contábeis necessárias para a elaboração dos índices, fato que ocorreu em quatro das 21 empresas insolventes utilizadas no estudo.

Para estudos futuros, sugere-se a aplicação de métodos de classificação diferentes, como a máquina de suporte vetorial, a análise por envoltória de dados e o *random forest*. Também podem ser utilizadas diferentes técnicas para a seleção das variáveis como as abordagens de filtro e *wrapper*, ao invés da análise fatorial, que por sua vez, também pode ser utilizada de forma diferente, sendo considerado os escores obtidos pelas empresas em cada dimensão, e não o valor da variável substituta.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALTMAN, Edward I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Backruptancy. **Journal of Finance**, Boston, v.23, p. 586 - 609. 1968.

ALTMAN, E. I.; BAIDYA, T. K. N.; DIAS, L. M. R. Previsão de problemas financeiros em empresas. **Revista de Administração de Empresas**, Rio de Janeiro, v. 19, n. 1, p. 17-28, 1979.

AGARWAL, Vineet; TAFFLER, Richard. Comparing the performance of market-based and accounting-based bankruptcy pediction models. **Journal of Banking & Finance**, v.32, n. 8, p. 1541-1551, 2008.

AITA, Jaqueline; ZANI, João; SILVA, Carlos E. S. Determinantes de insolvência bancária no Brasil: identificação de evidências macro e microeconômicas. *In*: XI Encontro Brasileiro de Finanças, 2011, Rio de Janeiro. **Anais...** Rio de Janeiro: FGV, 2011.

BARROS, Marcelo P. F. Lojas Arapuã e a globalização: a melhor do setor pede concordata. **Gestão e Planejamento**, Salvador, v. 1, n. 2, 2007

BASGALUPP, Márcio Porto. **LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão**. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Universidade de São Paulo, São Paulo, 2010.

BEAVER, William H. Financial Ratios as Predictors of Failure. **Journal of Accounting Research**, Chicago, Suplement of Selected Studies, p. 77-111. 1966.

BELLOVARY, Jodi; GIACOMINO, Don; AKERS, Michael. A review of bankruptcy prediction studies: 1930 to present. **Journal of Financial Education**, Milwaukee, v. 33, 2007.

BRITO, Giovani A. S.; ASSAF NET, Alexandre. Modelo de classificação de risco de crédito de empresas. **Revista de Contabilidade e Finanças**, São Paulo, v. 19, n. 46, p.18-29, 2008.

BRITO, Giovani A. S.; ASSAF NET, A.; CORRAR, L. J. Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil. **Revista de Contabilidade e Finanças**, São Paulo, v. 20, n. 51, p.28-43, 2009.

CARVALHO, Flávio Leonel et al. Identificação de indicadores contábeis relevantes para previsão e projeção de rentabilidade. **Revista de Educação e Pesquisa em Contabilidade**, Brasília, v. 4, n. 3, p. 94-110, 2010.

CASTRO JÚNIOR, F. H. F. **Previsão de insolvência de empresas brasileiras usando análise discriminante, regressão logística e redes neurais**. Dissertação (Mestrado em Administração) – Universidade de São Paulo, São Paulo, 2003.

CHUNG, Kim-Choy; TAN, S. S.; HOLDSWORTH, D. K. Insolvency prediction model using multivariate discriminante analysis and artificial neural network for finance industry in New Zeland. **International Journal of Business and Management**, Toronto, v. 3, n. 1, 2008.

CLARO, Carlos Roberto. **Recuperação Judicial: Sustentabilidade e função social da empresa**. Dissertação (Mestrado em Direito Empresarial e Cidadania), Centro Universitário de Curitiba, Curitiba, 2008.

CORRAR, Luiz J.; PAULO, Edilson; DIAS FILHO, José Maria. **Análise Multivariada para os cursos de Administração, Ciências Contábeis e Economia**. São Paulo: Atlas, 2012.

ELISABETSKY, Roberto. **Um modelo matemático para decisões de crédito no Banco Comercial**. 1976. Dissertação (Mestrado) – Instituto Politécnico da Universidade de São Paulo, 1976.

FITZPATRICK, Paul J. **A Comparison of the Ratios of the Successful Industrial Enterprises with those of Failed Companies**. The Accountants Publishing Company, 1932.

FLACH, Peter A. ROC Analysis. **Encyclopedia of Machine Learning**. Nova Iorque: Springer, 2010, 1031 p.

GUIMARÃES, Ailton; MOREIRA, Tito B. S. Previsão de insolvência: um modelo baseado em índices contábeis com utilização da análise discriminante. **Revista de Economia Contemporânea**, Rio de Janeiro. v. 12, n. 1, p. 151-178, jan./abr. 2008.

HAIR JUNIOR, J. F. et al. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005.

HAYKIN, Simon. **Redes Neurais: Princípios e prática**. Porto Alegre: Bookman, 2001.

HORTA, Rui A. Mathiasi. **Uma metodologia de mineração de dados para a previsão de insolvência de empresas brasileiras de capital aberto**. Tese (Doutorado em Engenharia Civil) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

HORTA, Rui. A. Mathiasi. *et al.* Comparação de técnicas de seleção de atributos para previsão de insolvência de empresas brasileiras no período 2005-2007. *In*: Encontro Nacional de Pós-Graduação e Pesquisa em Administração, 34, 2010, Rio de Janeiro. **Anais...** Rio de Janeiro: Anpad, 2010.

KAISER, Henry F. An index of factorial simplicity. **Psychometrika**, v. 39, n. 1, p 31-36, 1974

KANITZ, Stephen Charles. **Como prever falências**. São Paulo, Mc Graw-Hill, 1978.

KAWAGUCHI, Kiyoshi. **A multithreaded software model for backpropagation neural network applications**. Dissertação (Mestrado em Ciências), Universidade do Texas, El Paso, 2000. Disponível em: < http://www.ece.utep.edu/research/webfuzzy/docs/kk-thesis/kk-thesis-html/thesis.html>. Acesso em 10 nov 2013.

LACHTERMACHER, Gerson; ESPENCHITT, Dilson G. Previsão de falência de empresas: estudo de generalização de redes neurais. *In*: Encontro Nacional de Pós-Graduação e Pesquisa em Administração, 25, 2001, Campinas. **Anais...** São Paulo: Anpad, 2001.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 5 ed. São Paulo: Atlas, 2003.

LANDEIRO, Victor L. **Introdução ao uso do programa R**. Instituto Nacional de Pesquisa da Amazônia, Manaus, 2013. Disponível em: < http://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf>. Acesso em 15 out 2013.

MATIAS, Alberto B. **Indicadores contábeis e financeiros de previsão de insolvência: a experiência da pequena e média empresa**. Tese (Livre-Docência) – Universidade de São Paulo, São Paulo, 1976.

MATIAS, Alberto B.; SIQUEIRA, José de O.; Risco bancário: modelo de previsão de insolvência de bancos no Brasil. **Revista de Administração**, v. 31, p. 19-28, São Paulo, 1996.

MINUSSI, João A.; DAMASCENA, Cláudio; NESS JR, Walter L. Um modelo de previsão de solvência utilizando regressão logística. **Revista de Administração Contemporânea**, Curitiba. v. 6 n. 3 p. 109-128, set./dez. 2002.

NGUYEN, Huong Giang. Using Neural Network in Predicting Corporate Failure. **Journal of Social Sciences**, Sydney, v. 1, n. 4, 2005.

ODOM, M. D. SHARDA, R. *A neural network model for bankruptcy prediction*. **International Joint Conference on Neural Network**, v. 2, p. 163-168, 2010.

OHLSON, James A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. **Journal of Accounting Research**, Chicago, v. 18, n. 1, 1980.

ONUSIC, Luciana Massaro, et al; Estudo exploratório utilizando as técnicas de análise por envoltória de dados e redes neurais artificiais na previsão de insolvência de empresas. **FACEF Pesquisa**, v. 9, n. 2, 2006.

ONUSIC, Luciana M.; KASSAI, Silvia; VIANA, Adriana B. N. Comparação dos resultados de utilização de análise por envoltória de dados e regressão logística em modelos de previsão de insolvência: um estudo aplicado a empresas brasileiras. **FACEF Pesquisa**, v. 7, n. 1, 2004.

PEDRO, Sílvia M. D. **Exploração de dados aplicada à análise de risco de crédito**. Monografia (Licenciatura em Matemática Aplicada e Computação) – Instituto Superior Técnico, Lisboa, 2001.

PEREIRA, Orlando M.; NESS JÚNIOR, Walter Lee. O modelo E-Score de previsão de falências para empresas de Internet. **Revista de Administração Contemporânea**, Curitiba, v. 8, n. 3, p. 143-166, 2004.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, 2008.

R Core Team. **R: A language and enviroment for statistical computing**. R Foundation for Statistical Computing, Viena, Áustria, 2013.

REBOUÇAS, S. M. D. P. **Metodologias de classificação supervisionada para análise de dados de microarrays**. Tese (Doutorado em Estatística e Investigação Operacional) – Universidade de Lisboa, Lisboa, 2011.

REIS, Edna A.; REIS, Ilka A. **Análise descritiva de dados: síntese numérica**. 1. ed. Belo Horizonte: Universidade Federal de Minas Gerais, 2002. 36p.

SANVICENTE, Antônio Z.; MINARDI, Andrea M. A. F.; Identificação de indicadores contábeis significativos para previsão de concordata de empresas. **Finance Lab Working Papers**, IBMEC, São Paulo, 1998.

SANTOS, José Odálio dos. Análise comparativa de métodos para previsão de insolvência em uma carteira de crédito bancário de empresas de médio porte. **Revista de Gestão USP**, São Paulo, v. 15, n. 3, p.11-24, 2008.

SANTOS, M. F. *et al*. Corporate bankruptcy prediction using data mining techniques. **Data Mining VII: Data, Text and Web Mining and their Business Applications**, Ashurst Lodge: WIT Press, v. 37, 2006.

SCARPEL, Rodrigo A. Previsão de insolvência de empresas utilizando support vector machine. **Revista de Economia e Administração**, São Paulo, v. 7, n. 3, p. 281-295, 2008.

SILVA, Rômulo O. R. C. **Estudo de insolvência de empresas de capital aberto**. Trabalho acadêmico (Mestrado em Administração) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2006.

SILVA, José Pereira da. **Modelos para classificação de empresas com vistas à concessão de crédito**. Dissertação (Mestrado em Administração) – Escola de Administração de Empresas de São Paulo, Fundação Getúlio Vargas, São Paulo, 1982.

TIMOFEEV, Roman. **Classification and Regression Trees (CART) Theory and Appications**. Dissertação (Mestrado em Artes) – Universidade de Humboldt, Berlin, 2004.

VIRGILLITO, Salvatore B.; FAMÁ, Rubens. Estatística multivariada na construção de modelos para análise do risco de crédito e previsão de insolvência de empresas. **Revista Integração**, São Paulo, n. 53, p. 105-118, 2008.

YIM, Juliana; MITCHEL, Heather. A comparison of corporate distress prediction models in Brazil: hydrid neural networks, logit models and discriminant analysis. **Nova Economia**, Belo Horizonte, v. 15, n. 1, p. 73-93, 2005.

APÊNDICES

APÊNDICE A – Empresas utilizadas no estudo

Índice	Nome	Ano	Setor	Situação
1	Agrenco	2008	Agroindústria	Recuperação Judicial
2	Lark Maqs	2012	Comércio de máquinas e veículos pesados	Recuperação Judicial
3	Const Beter	2008	Construção e engenharia	Recuperação Judicial
4	Tecnosolo	2012	Construção e engenharia	Recuperação Judicial
5	IGB S/A	2010	Eletroeletrônicos	Recuperação Judicial
6	Celpa	2012	Energia elétrica	Recuperação Judicial
7	Rede Energia	2012	Energia elétrica	Recuperação Judicial
8	Eucatex	2003	Madeira e papel	Recuperação Judicial
9	Chiarelli	2009	Material de construção	Recuperação Judicial
10	Recrusul	2006	Material de transporte	Recuperação Judicial
11	Kepler Weber	2007	Metalurgia	Recuperação Judicial
12	Sansuy	2005	Químico	Recuperação Judicial
13	F Guimaraes	2007	Tecidos, vestuário e calçados	Recuperação Judicial
14	Botucatu Tex	2008	Tecidos, vestuário e calçados	Recuperação Judicial
15	Tecel S Jose	2010	Tecidos, vestuário e calçados	Recuperação Judicial
16	Tex Renaux	2010	Tecidos, vestuário e calçados	Recuperação Judicial
17	Buettner	2011	Tecidos, vestuário e calçados	Recuperação Judicial
18	Fab C Renaux	2011	Tecidos, vestuário e calçados	Recuperação Judicial
19	Schlosser	2011	Tecidos, vestuário e calçados	Recuperação Judicial
20	Teka	2012	Tecidos, vestuário e calçados	Recuperação Judicial
21	Savarg	2005	Transportes aéreos	Recuperação Judicial
22	Rasip Agro	2008	Agroindústria	Normal
23	Renar	2008	Agroindústria	Normal
24	SLC Agricola	2008	Agroindústria	Normal
25	V-Agro	2008	Agroindústria	Normal
26	Minasmaquinas	2012	Comércio de máquinas e veículos pesados	Normal
27	Wlm Ind Com	2012	Comércio de máquinas e veículos pesados	Normal
28	Azevedo	2008	Construção e engenharia	Normal
29	Lix da Cunha	2008	Construção e engenharia	Normal
30	Sultepa	2008	Construção e engenharia	Normal
31	Cr2	2012	Construção e engenharia	Normal
32	Mills	2012	Construção e engenharia	Normal
33	Sondotecnica	2012	Construção e engenharia	Normal
34	Trisul	2012	Construção e engenharia	Normal
35	Bematech	2010	Eletroeletrônicos	Normal
36	Itautec	2010	Eletroeletrônicos	Normal
37	Positivo Inf	2010	Eletroeletrônicos	Normal

38	Springer	2010	Eletroeletrônicos	Normal
39	Whirlpool	2010	Eletroeletrônicos	Normal
40	AES Elpa	2012	Energia elétrica	Normal
41	Celesc	2012	Energia elétrica	Normal
42	Celpe	2012	Energia elétrica	Normal
43	Cemar	2012	Energia elétrica	Normal
44	Cemat	2012	Energia elétrica	Normal
45	Coelce	2012	Energia elétrica	Normal
46	Eletropaulo	2012	Energia elétrica	Normal
47	Energias BR	2012	Energia elétrica	Normal
48	Light S/A	2012	Energia elétrica	Normal
49	Tractebel	2012	Energia elétrica	Normal
50	Celul Irani	2003	Madeira e papel	Normal
51	Duratex	2003	Madeira e papel	Normal
52	Fibria	2003	Madeira e papel	Normal
53	Klabin S/A	2003	Madeira e papel	Normal
54	Eternit	2009	Material de construção	Normal
55	Fras-Le	2006	Material de transporte	Normal
56	Plascar Part	2006	Material de transporte	Normal
57	Riosulense	2006	Material de transporte	Normal
58	Wetzel S/A	2006	Material de transporte	Normal
59	Wiest	2006	Material de transporte	Normal
60	Aliperti	2007	Metalurgia	Normal
61	Forja Taurus	2007	Metalurgia	Normal
62	Lupatech	2007	Metalurgia	Normal
63	Mangels Indl	2007	Metalurgia	Normal
64	Dixie Toga	2005	Químico	Normal
65	Elekeiroz	2005	Químico	Normal
66	Evora	2005	Químico	Normal
67	Pronor	2005	Químico	Normal
68	Santanense	2007	Tecidos, vestuário e calçados	Normal
69	Vulcabras	2007	Tecidos, vestuário e calçados	Normal
70	Guararapes	2008	Tecidos, vestuário e calçados	Normal
71	Karsten	2008	Tecidos, vestuário e calçados	Normal
72	Alpargatas	2010	Tecidos, vestuário e calçados	Normal
73	Cambuci	2010	Tecidos, vestuário e calçados	Normal
74	Le Lis Blanc	2010	Tecidos, vestuário e calçados	Normal
75	Arezzo Co	2011	Tecidos, vestuário e calçados	Normal
76	Cedro	2011	Tecidos, vestuário e calçados	Normal
77	Cremer	2011	Tecidos, vestuário e calçados	Normal
78	Dohler	2011	Tecidos, vestuário e calçados	Normal
79	Ind Cataguas	2011	Tecidos, vestuário e calçados	Normal
80	Pettenati	2011	Tecidos, vestuário e calçados	Normal
81	Vicunha Text	2011	Tecidos, vestuário e calçados	Normal
82	Wembley	2011	Tecidos, vestuário e calçados	Normal
83	Coteminas	2012	Tecidos, vestuário e calçados	Normal

84	Marisol	2012	Tecidos, vestuário e calçados	Normal
85	Springs	2012	Tecidos, vestuário e calçados	Normal
86	Gol	2005	Transportes aéreos	Normal
87	Tam S/A	2005	Transportes aéreos	Normal

APÊNDICE B – Classificação detalhada dos modelos.

Índice da Empresa	Situação Real	LDA	RL	kNN	CART	ANN
1	Insolvente	Solvente	Insolvente	Solvente	Solvente	Insolvente
2	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
3	Insolvente	Insolvente	Insolvente	Solvente	Insolvente	Insolvente
4	Insolvente	Solvente	Solvente	Solvente	Solvente	Insolvente
5	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
6	Insolvente	Solvente	Insolvente	Insolvente	Insolvente	Insolvente
7	Insolvente	Solvente	Insolvente	Solvente	Insolvente	Insolvente
8	Insolvente	Solvente	Solvente	Insolvente	Insolvente	Insolvente
9	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
10	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
11	Insolvente	Insolvente	Insolvente	Solvente	Insolvente	Insolvente
12	Insolvente	Solvente	Solvente	Solvente	Solvente	Insolvente
13	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
14	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
15	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
16	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
17	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
18	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
19	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
20	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
21	Insolvente	Insolvente	Insolvente	Insolvente	Solvente	Insolvente
22	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
23	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
24	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
25	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
26	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
27	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
28	Solvente	Solvente	Insolvente	Solvente	Solvente	Solvente
29	Solvente	Solvente	Insolvente	Solvente	Solvente	Solvente
30	Solvente	Solvente	Solvente	Insolvente	Solvente	Solvente
31	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
32	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
33	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
34	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
35	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
36	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
37	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
38	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
39	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente

40	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
41	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
42	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
43	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
44	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
45	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
46	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
47	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
48	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
49	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
50	Solvente	Solvente	Solvente	Insolvente	Solvente	Solvente
51	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
52	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
53	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
54	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
55	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
56	Solvente	Solvente	Insolvente	Solvente	Solvente	Solvente
57	Solvente	Solvente	Solvente	Insolvente	Solvente	Solvente
58	Solvente	Solvente	Insolvente	Insolvente	Solvente	Solvente
59	Solvente	Insolvente	Insolvente	Insolvente	Insolvente	Insolvente
60	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
61	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
62	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
63	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
64	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
65	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
66	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
67	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
68	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
69	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
70	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
71	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
72	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
73	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
74	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
75	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
76	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
77	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
78	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
79	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
80	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
81	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
82	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
83	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
84	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
85	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
86	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente
87	Solvente	Solvente	Solvente	Solvente	Solvente	Solvente

APÊNDICE C – Matrizes de correlação antes e depois da Análise Fatorial

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
X_1	1,00	0,50	0,46	0,04	-0,90	0,47	0,77	0,45	0,45	0,26	0,25	0,30	0,23	-0,14	0,12	0,12
X_2	0,50	1,00	0,30	0,37	-0,37	0,37	0,81	0,41	0,30	0,72	0,15	0,83	0,14	-0,08	0,08	0,24
X_3	0,46	0,30	1,00	0,07	-0,56	0,76	0,51	0,74	0,82	0,32	0,31	0,40	0,54	0,07	0,11	0,01
X_4	0,04	0,37	0,07	1,00	0,02	0,15	0,25	0,19	0,22	0,41	0,24	0,44	-0,01	0,20	-0,02	-0,07
X_5	-0,90	-0,37	-0,56	0,02	1,00	-0,47	-0,73	-0,45	-0,49	-0,25	-0,53	-0,28	-0,22	-0,09	-0,09	-0,21
X_6	0,47	0,37	0,76	0,15	-0,47	1,00	0,51	0,93	0,76	0,35	0,28	0,40	0,74	-0,01	-0,02	-0,08
X_7	0,77	0,81	0,51	0,25	-0,73	0,51	1,00	0,52	0,48	0,52	0,28	0,60	0,23	-0,01	0,11	0,24
X_8	0,45	0,41	0,74	0,19	-0,45	0,93	0,52	1,00	0,76	0,46	0,28	0,49	0,77	-0,01	-0,01	-0,08
X_9	0,45	0,30	0,82	0,22	-0,49	0,76	0,48	0,76	1,00	0,33	0,29	0,41	0,46	0,19	0,31	-0,10
X_{10}	0,26	0,72	0,32	0,41	-0,25	0,35	0,52	0,46	0,33	1,00	0,18	0,83	0,22	-0,09	0,12	0,07
X_{11}	0,25	0,15	0,31	0,24	-0,53	0,28	0,52	0,28	0,29	0,18	1,00	0,20	0,12	-0,05	0,06	0,15
X_{12}	0,30	0,83	0,40	0,44	-0,28	0,40	0,60	0,49	0,41	0,83	0,20	1,00	0,24	-0,05	0,15	0,10
X_{13}	0,23	0,14	0,54	-0,01	-0,22	0,74	0,23	0,77	0,46	0,22	0,12	0,24	1,00	-0,01	0,27	-0,03
X_{14}	-0,14	-0,08	0,07	0,20	0,12	0,26	-0,01	-0,01	0,19	-0,09	0,12	-0,05	-0,01	1,00	0,01	-0,12
X_{15}	0,12	0,08	0,11	-0,02	-0,09	0,21	0,11	0,27	0,12	0,12	0,06	0,15	0,27	0,01	1,00	-0,11
X_{16}	0,12	0,24	0,01	-0,07	-0,21	-0,08	0,24	-0,08	-0,10	0,07	0,15	0,10	-0,03	-0,12	-0,11	1,00

	X1	X13	X3	X11	X14	X16	X15	X12
X1	1,00	0,23	0,46	0,25	-0,14	0,12	0,12	0,30
X13	0,23	1,00	0,54	0,12	-0,01	-0,03	0,27	0,24
X3	0,46	0,54	1,00	0,31	0,07	0,01	0,11	0,40
X11	0,25	0,12	0,31	1,00	0,12	0,15	0,06	0,20
X14	-0,14	-0,01	0,07	0,12	1,00	-0,12	0,01	-0,05
X16	0,12	-0,03	0,01	0,15	-0,12	1,00	-0,11	0,10
X15	0,12	0,27	0,11	0,06	0,01	-0,11	1,00	0,15
X12	0,30	0,24	0,40	0,20	-0,05	0,10	0,15	1,00

APÊNDICE D – Probabilidades utilizadas para as curvas ROC

Índice	LDA	GLM	kNN	CART	ANN
1	0,625070	0,624978	1,000000	0,967742	0,000000
2	0,264827	0,110926	0,000000	0,055556	0,000000
3	0,113491	0,042686	1,000000	0,055556	0,000000
4	0,913175	0,967454	1,000000	0,967742	0,055770
5	0,291781	0,092127	0,000000	0,055556	0,000000
6	0,615641	0,568860	0,000000	0,055556	0,000000
7	0,671414	0,638140	1,000000	0,055556	0,000000
8	0,840163	0,901103	0,000000	0,055556	0,000227
9	0,173686	0,041391	0,000000	0,055556	0,000000
10	0,171724	0,045310	0,000000	0,055556	0,000000
11	0,162443	0,048850	1,000000	0,055556	0,111451
12	0,892027	0,950562	1,000000	0,714286	0,000000
13	0,205989	0,048294	0,000000	0,055556	0,000000
14	0,471909	0,389093	0,000000	0,055556	0,000002
15	0,178859	0,044808	0,000000	0,055556	0,000000
16	0,246588	0,081941	0,000000	0,055556	0,000000
17	0,470132	0,428089	0,000000	0,055556	0,000000
18	0,364969	0,194344	0,000000	0,055556	0,000000
19	0,327529	0,160104	0,000000	0,055556	0,000000
20	0,275463	0,096341	0,000000	0,055556	0,114706
21	0,345159	0,136780	0,000000	0,714286	0,000000
22	0,953744	0,989136	1,000000	0,967742	1,000000
23	0,830491	0,945338	1,000000	0,967742	1,000000
24	0,955305	0,992835	1,000000	0,967742	1,000000
25	0,766071	0,875031	1,000000	0,967742	0,999987
26	0,884962	0,939149	1,000000	0,967742	1,000000
27	0,999997	1,000000	1,000000	0,967742	1,000000
28	0,583178	0,460069	1,000000	0,967742	1,000000
29	0,594303	0,557702	1,000000	0,967742	0,999891
30	0,687831	0,711969	0,000000	0,714286	1,000000
31	0,914897	0,973317	1,000000	0,967742	1,000000
32	0,977564	0,996199	1,000000	0,967742	1,000000
33	0,997153	0,999900	1,000000	0,967742	1,000000
34	0,685253	0,702380	1,000000	0,967742	1,000000
35	0,999354	0,999987	1,000000	0,967742	1,000000
36	0,819618	0,882203	1,000000	0,967742	1,000000
37	0,927166	0,972774	1,000000	0,967742	1,000000
38	0,946635	0,993194	1,000000	0,967742	1,000000
39	0,889579	0,942271	1,000000	0,967742	1,000000
40	0,913702	0,949036	1,000000	0,967742	1,000000
41	0,870542	0,929730	1,000000	0,967742	0,999681

42	0,887820	0,941774	1,000000	0,967742	1,000000
43	0,855912	0,906309	1,000000	0,967742	1,000000
44	0,870779	0,924044	1,000000	0,967742	0,999714
45	0,956395	0,986003	1,000000	0,967742	1,000000
46	0,946441	0,978768	1,000000	0,967742	0,999420
47	0,881702	0,936797	1,000000	0,967742	1,000000
48	0,792969	0,836670	1,000000	0,967742	1,000000
49	0,960536	0,987859	1,000000	0,967742	1,000000
50	0,907075	0,957189	0,000000	0,714286	1,000000
51	0,954003	0,988588	1,000000	0,967742	1,000000
52	0,835120	0,895993	1,000000	0,967742	1,000000
53	0,818171	0,853110	1,000000	0,714286	1,000000
54	0,999697	0,999979	1,000000	0,967742	1,000000
55	0,955369	0,986453	1,000000	0,967742	0,999911
56	0,631116	0,582400	1,000000	0,714286	0,999941
57	0,787071	0,796490	0,000000	0,714286	1,000000
58	0,545487	0,381670	0,000000	0,967742	0,999927
59	0,154755	0,024757	0,000000	0,055556	0,000000
60	0,878263	0,964987	1,000000	0,967742	1,000000
61	0,941150	0,982519	1,000000	0,967742	1,000000
62	0,877729	0,938564	1,000000	0,967742	1,000000
63	0,727098	0,759294	1,000000	0,967742	1,000000
64	0,940065	0,974395	1,000000	0,967742	0,994249
65	0,993860	0,999323	1,000000	0,967742	1,000000
66	0,974546	0,995662	1,000000	0,967742	1,000000
67	0,968555	0,992003	1,000000	0,967742	1,000000
68	0,947451	0,987150	1,000000	0,967742	1,000000
69	0,942652	0,979075	1,000000	0,967742	1,000000
70	0,962297	0,995574	1,000000	0,967742	1,000000
71	0,911557	0,964088	1,000000	0,967742	1,000000
72	0,984208	0,998057	1,000000	0,967742	1,000000
73	0,972911	0,988631	1,000000	0,967742	1,000000
74	0,997360	0,999906	1,000000	0,967742	1,000000
75	0,998313	0,999805	1,000000	0,967742	0,985419
76	0,890531	0,954082	1,000000	0,967742	1,000000
77	0,946799	0,985893	1,000000	0,967742	1,000000
78	0,996189	0,999893	1,000000	0,967742	1,000000
79	0,914078	0,971124	1,000000	0,967742	0,999454
80	0,749360	0,814916	1,000000	0,967742	1,000000
81	0,948971	0,987134	1,000000	0,967742	1,000000
82	0,703653	0,719879	1,000000	0,967742	1,000000
83	0,815289	0,886340	1,000000	0,967742	0,999996
84	0,989988	0,999194	1,000000	0,967742	1,000000
85	0,831498	0,905788	1,000000	0,967742	0,999790
86	0,999449	0,999967	1,000000	0,967742	1,000000
87	0,853092	0,873349	1,000000	0,967742	1,000000

APÊNDICE E – Códigos utilizados no R

```
####### Importação dos dados
dados<-
read.table('C:/Users/Romulo/Dropbox/Monografia/Dados/Variáveis.txt',header=
T,dec=",")
dados2<-dados[,-1] # Cria uma matriz de dados sem a coluna de classes
attach(dados)
####### Análise dos dados
summary(dados2) # Estatísticas descritivas das variáveis
matcor<-cor(dados2) # Cria a matriz de correlação das variáveis
write.table(matcor,"matcor.csv")
####### Análise Fatorial
cortest.bartlett(dados2) # Teste de Esfericidade de Barttlet
# Código para obtenção do KMO criado por Trujillo-Oritz
kmo = function( data ){
  library(MASS)
  X <- cor(as.matrix(data))
  iX <- ginv(X)
  S2 <- diag(diag((iX^-1)))
  AIS <- S2%*%iX%*%S2                      # anti-image covariance matrix
  IS <- X+AIS-2*S2                         # image covariance matrix
  Dai <- sqrt(diag(diag(AIS)))
  IR <- ginv(Dai)%*%IS%*%ginv(Dai)         # image correlation matrix
  AIR <- ginv(Dai)%*%AIS%*%ginv(Dai)       # anti-image correlation matrix
  a <- apply((AIR - diag(diag(AIR)))^2, 2, sum)
  AA <- sum(a)
  b <- apply((X - diag(nrow(X)))^2, 2, sum)
  BB <- sum(b)
  MSA <- b/(b+a)                        # indiv. measures of sampling adequacy
  AIR <- AIR-diag(nrow(AIR))+diag(MSA)  # Examine the anti-image of the
                                        # correlation matrix. That is the
                                        #    negative    of    the    partial
correlations,
                                        #   partialling   out   all   other
variables.

  kmo <- BB/(AA+BB)                     # overall KMO statistic
  # Reporting the conclusion
    if (kmo >= 0.00 && kmo < 0.50){
```

```
      test <- 'The KMO test yields a degree of common variance
unacceptable for FA.'
    } else if (kmo >= 0.50 && kmo < 0.60){
      test <- 'The KMO test yields a degree of common variance miserable.'
    } else if (kmo >= 0.60 && kmo < 0.70){
      test <- 'The KMO test yields a degree of common variance mediocre.'
    } else if (kmo >= 0.70 && kmo < 0.80){
      test <- 'The KMO test yields a degree of common variance middling.'
    } else if (kmo >= 0.80 && kmo < 0.90){
      test <- 'The KMO test yields a degree of common variance meritorious.'
    } else {
      test <- 'The KMO test yields a degree of common variance marvelous.'
    }
    ans <- list(  overall = kmo,
                  report = test,
                  individual = MSA,
                  AIS = AIS,
                  AIR = AIR )
    return(ans)
}

kmo(dados2)$overall

# Obtenção dos fatores
fit<-principal(dados2, nfactors=8, rotate="varimax",scores=TRUE)
load<-fit$loadings
load
write.table(load,'escores.csv')
dados3<-data.frame(Y,X1,X3,X11,X12,X13,X14,X16,X15) # Monta a matriz com as
# variáveis selecionadas
dados4<-dados3[,-1] # Monta a matriz sem a coluna das classes
######## Análise Discriminante Linear
# Teste de Normalidade
dados5<-cbind(X1,X13,X3,X11,X14,X16,X15,X12)
mshapiro.test(t(dados5[1:21,]))
mshapiro.test(t(dados5[22:87,]))
kms<-numeric(0)
for (i in 1:ncol(dados4))
{
 kms<-shapiro.test(dados4[1:22,i])
 print(colnames(dados4[i]))
```

```
 print(kms)
}
kms2<-numeric(0)
for (i in 1:ncol(dados4))
{
 kms2<-shapiro.test(dados4[23:92,i])
 print(colnames(dados4[i]))
 print(kms2)
}
# Teste da igualdade das variâncias
fligner.test(dados4,Y)
fligner.test(X1,Y)
fligner.test(X3,Y)
fligner.test(X11,Y)
fligner.test(X12,Y)
fligner.test(X13,Y)
fligner.test(X14,Y)
fligner.test(X15,Y)
fligner.test(X16,Y)
# Aplicação da regressão stewise e criação do modelo
linear<-
stepclass(y,data=dados3,method="lda",improvement=0.001,fold=87,direction="b
oth")
modelo.lda<-lda(linear$formula,data=dados3)
# Validação cruzada
classif.lda<-data.frame(0)
predicao.lda<-data.frame(0)
valor<-numeric(0)
for(i in 1:nrow(dados4))
{
 modelo.lda<-lda(linear$formula, data=dados[-i,])
 predict.lda<-predict(modelo.lda,newdata=dados[i,])
 valor[i]<-predict.lda$class
 classif.lda[i,]<-valor[i]-1
 predicao.lda[i,]<-predict.lda$posterior[,2]
}
####### Regressão Logística
# Seleção das variáveis e criação do modelo
modelo.glm<-glm(y, data=dados3, family=binomial(link=logit))
stepwise(modelo.glm)
modelo.glm<-glm(Y ~ X3+X12, data=dados4,family=binomial(link=logit))
```

```
# Validação cruzada
classif.glm<-data.frame(0)
valor2<-numeric(0)
for ( i in 1:nrow(dados4) )
{
 modelo.glm<-glm(Y ~ X3+X12, data=dados[-i,],family=binomial(link=logit))
 predict.glm<-predict(modelo.glm,newdata=dados[i,],type="response")
 valor2[i]<-predict.glm
 predicao.glm[i,]<-valor2[i]
}
# Teste de significância das variáveis e coeficientes
anova(modelo.glm,test="Chisq")
summary(modelo.glm,correlation=T)
vif(modelo.glm)
####### Vizinhos Mais Próximos
# Obtenção do número ideal de vizinhos
K<-numeric(0)
for (i in 1:1000)
{
 tune<-tune.knn(dados4,as.factor(Y),k=1:10)
 K[i]<-tune$best.parameters$k
}
table(K)
# Criação do modelo com validação cruzada
predicao.knn<-data.frame(0)
probabilidade.knn<-data.frame(0)
for (i in 1:nrow(dados4))
{
 pred<-knn(dados4[-i,],dados4[i,],Y[-i],prob=T,k=1,l=0,use.all=T)
 probabilidade.knn[i]<-attr(pred,"prob")
 predicao.knn[i]<-pred
}
####### Árvores de Classificação
# Criação da árvore
cart<-rpart(y,data=dados4,method="class")
summary(cart)
# Validação cruzada
predicao.cart<-data.frame(0)
for(i in 1:nrow(dados4))
{
```

```
 modelo.cart<-rpart(y, data=dados[-i,],method="class")
 classif.cart<-predict(modelo.cart,data=dados[i,],method="class")
 predicao.cart<-(classif.cart[,2])
}
####### Redes Neurais Artificiais
tamanho<-numeric(0)
for(i in 1:1000)
{
 tamanho[i]<-tune.nnet(y, data=dados3,size=1:5)$best.parameters$size
}
table(tamanho)
# Criação da rede
rna<-nnetrandom(y, data=dados3,size=5, tries=30, skip=F, rang=0.00001,
maxit=10000, na.action=na.exclude)
# Validação cruzada
predicao.ann<-numeric(0)
for(i in 1:nrows(dados4))
{
  rna<-nnetrandom(y, dados4[-i,],size=5, tries=30, skip=F, rang=0.00001,
maxit=10000, na.action=exclude)
 predicao.ann[i]<-predict(rna,dados4[i,])
}
####### Curvas ROC e AUC
roc.lda<-prediction(predicao.lda,Y)
roc.glm<-prediction(predicao.glm,Y)
roc.cart<-prediction(predicao.cart,Y)
roc.knn<-prediction(predicao.knn,Y)
roc.ann<-prediction(predicao.ann,Y)
perf.lda<-performance(roc.lda,measure="tpr",x.measure="fpr")
perf.glm<-performance(roc.glm,measure="tpr",x.measure="fpr")
perf.knn<-performance(roc.knn,measure="tpr",x.measure="fpr")
perf.cart<-performance(roc.cart,measure="tpr",x.measure="fpr")
perf.ann<-performance(roc.ann,measure="tpr",x.measure="fpr")
auc.lda<-performance(roc.lda,measure="auc")
auc.glm<-performance(roc.glm,measure="auc")
auc.knn<-performance(roc.knn,measure="auc")
auc.cart<-performance(roc.cart,measure="auc")
auc.ann<-performance(roc.ann,measure="auc")
```