UNIVERSIDADE FEDERAL DO PARANÁ

DIOGO CASSIN DE CARVALHO OLIVEIRA

MÉTODOS MULTIVARIADOS APLICADOS À ANÁLISE FINANCEIRA DE EMPRESAS DO SETOR AUTOMOTIVO

CURITIBA

2016

DIOGO CASSIN DE CARVALHO OLIVEIRA

MÉTODOS MULTIVARIADOS APLICADOS À ANÁLISE FINANCEIRA DE EMPRESAS DO SETOR AUTOMOTIVO

Dissertação apresentada ao Curso de Pós-Graduação em Engenharia de Produção, Área de Concentração em Pesquisa Operacional, Linha de Pesquisa em Métodos Estatísticos Aplicados à Engenharia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Engenharia de Produção.

Orientadora: Profa. Dra. Sonia Isoldi Marty Gama Müller

CURITIBA

2016

O482 Oliveira, Diogo Cassin de Carvalho.
Métodos multivariados aplicados à análise financeira de empresas do setor automotivo. / Diogo Cassin de Carvalho Oliveira. – Curitiba, 2016.
143 f. : il. color.; 30cm.

Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-graduação em Engenharia de Produção.

Orientadora : Profa. Dra. Sonia Isoldi Marty Gama Müller.

1. Engenharia de produção. 2. Indústria automobilística. 3. Análise econômico-financeira. I. Müller, Sonia Isoldi Marty Gama. II. Título.

CDU 657.3

TERMO DE APROVAÇÃO

DIOGO CASSIN DE CARVALHO OLIVEIRA

MÉTODOS MULTIVARIADOS APLICADOS À ANÁLISE FINANCEIRA DE EMPRESAS DO SETOR AUTOMOTIVO

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Curso de Pós-Graduação em Engenharia de Produção, Setor de Tecnologia da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientadora: Profa. Dra. Sonia Isoldi Marty Gama Müller
Programa de Pós-Graduação em Engenharia de Produção - UFPR.

Prof. Dr. Marcelo Gechele Cleto
Programa de Pós-Graduação em Engenharia de Produção - UFPR.

Profa. Dra. Neida Maria Patias Volpi
Programa de Pós-Graduação em Engenharia de Produção - UFPR.

Prof. Dr. Jair Mendes Marques
PPGMNE - UFPR.

Curitiba, 24 de novembro de 2016.

AGRADECIMENTOS

A Deus, por cuidar de mim e de minha família, concedendo-me a oportunidade de alcançar sonhos outrora muito distantes.

À minha esposa Samara, por ser minha companheira, motivar-me, apoiar-me incondicionalmente e estar presente em mais este momento importante de minha vida.

Aos meus pais João Batista de Oliveira e Dalva Cassin de Carvalho, por me ensinarem os valores da vida e por me incentivarem a buscar meus anseios.

A meus irmãos, minha sogra e demais familiares por todo o apoio.

À minha orientadora, Profa. Dra. Sonia I.M.G. Müller, por todos os ensinamentos e por toda a confiança depositada em nosso trabalho.

Ao Programa de Pós-Graduação em Engenharia de Produção da Universidade do Estado do Paraná, por todo o aprendizado e experiência obtida ao longo desta jornada.

Aos colegas do HSBC Seguros, por confiarem em meu trabalho e disponibilizar-me os horários necessários para cumprimento dos créditos do Programa de Mestrado.

“Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender e conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer”.

ALBERT EINSTEIN

RESUMO

A indústria automotiva exerce forte influência sobre a economia global e dispõe de uma larga variedade de companhias multinacionais, as quais possuem singularidades em termos de valores culturais, estratégia de negócios e políticas de investimentos. Neste contexto, esta pesquisa tem o objetivo de propor mecanismos cientificamente adequados para avaliação da rentabilidade de montadoras de veículos, como alternativa a processos dotados de subjetividade. A aplicação de métodos multivariados permite prever o nível de retorno gerado pelos negócios destas companhias, em função de índices financeiros derivados de suas demonstrações contábeis. Ademais, a aplicação destes modelos contribui para o enriquecimento da literatura acadêmica e visa subsidiar decisões de planejamento estratégico destas organizações, gerando benefícios a acionistas, clientes e fornecedores. De tal modo, foram coletadas informações financeiras de trinta montadoras de veículos com ações listadas em Bolsa de Valores, das quais foram selecionados índices financeiros de liquidez, endividamento, atividade e rentabilidade. A técnica de Análise de Agrupamentos foi empregada com o intuito de criar dois *clusters* para separar empresas de alta e baixa rentabilidade, alcançando resultados satisfatórios com formação de dois grupos distintos com baixa variância intragrupos e diferença significativa entre os agrupamentos. Após a definição dos grupos, as técnicas de Análise Discriminante de Fisher e Regressão Logística foram aplicadas de forma a identificar os índices financeiros necessários para predizer a rentabilidade de uma empresa e alocá-la em um dos grupos de rentabilidade. No que tange aos resultados alcançados neste trabalho, observou-se qualidade no ajustamento e alto poder preditivo de ambos os métodos, com elevado percentual de acertos. Igualmente, a área sob a curva ROC em ambos os métodos atingiu valores acima de 0,90, o que indica excelente desempenho ao classificar empresas. Após comparação entre os resultados gerados pelos dois métodos, selecionou-se o modelo logístico estimado para análise da rentabilidade de montadoras de veículos.

Palavras-chave: Índices financeiros. Métodos multivariados. Rentabilidade. Setor automotivo.

ABSTRACT

Automotive industry conspicuously drives the economy worldwide with a broad variety of global companies, which have singularities in terms of its culture, business strategy and investment decisions. Thereby, this research has been designed with the aim to propose scientifically suitable mechanisms to evaluate the profitability of automakers, in opposition to other processes spotted by subjectivity. The employment of multivariate methods provides the prediction of the yields added by the company's operating cash flow through financial ratios derived from its income statement and balance sheet. Moreover, the application of such models embeds value to the academic literature and aims to support decisions involving strategic planning of these companies, promoting benefits to shareholders, clients and suppliers. Thus, there have been collected financial ratios of thirty automakers listed on the stock market, such as liquidity, debt, activity and profitability. A Cluster Analysis has been performed with the aim to set two dissimilar groups to separate the companies with high or low profitability, achieving satisfactory results of building two clusters with low intern variance. Given that the groups have been properly set, there have been applied Fisher's Discriminant Analysis and Logistic Regression models, based on a training dataset, to identify the financial ratios which are capable to predict the company's profitability and allocate it into one of the groups according to its profitability. With regard to the results accomplished by this research, there have been observed quality of fitting and high predictive power of both models, with high correct classification rate. Moreover, the area under ROC curve reached values above 0.90, which indicates high quality of models' performance by classifying the companies. After comparing the results reached by the both methods, the estimated logistic model has been selected to evaluate the automakers' profitability.

Key-words: Automotive industry. Financial ratios. Multivariate methods. Profitability.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo dos anos, o setor automotivo brasileiro tem se consolidado como um pilar para a economia local, com relevante participação sobre o Produto Interno Bruto do país e geração de empregos diretos e indiretos. De forma geral, a robustez do setor automotivo configura-se como um indicador do progresso da economia brasileira, haja vista os benefícios gerados pela natureza multinacional de uma forte indústria automobilística no país.

De acordo com o Ministério da Indústria e Comercio Exterior (MDIC, 2015), a estratégia global de investimento da indústria automobilística em novas unidades produtivas envolve o fluxo de investimentos diretos externos para países e regiões em desenvolvimento com objetivo de deslocar a produção mundial e elevar a participação no mercado de países emergentes. Tal fato decorre da saturação do mercado nas nações desenvolvidas e da criação de blocos econômicos regionais, com livre comércio entre seus membros e incidência de barreiras comerciais às mercadorias de países externos ao grupo.

A indústria é composta por uma gama de companhias multinacionais, que operam em escala global ou em determinadas regiões, por meio de plantas. O setor engloba ampla variedade de empresas cuja atividade principal está relacionada com a fabricação, *design* e comercialização de autopeças ou veículos. Ademais, o mercado altamente competitivo demanda a modernização de processos e conhecimento dos fatores que exercem influência sobre o desempenho das empresas do segmento automotivo.

Pesquisadores acadêmicos como Kanitz (1976), demonstram interesse no desenvolvimento de modelos matemáticos para prever o desempenho de empresas e gerenciar os riscos financeiros incorridos em suas atividades. Tal abordagem vislumbra a possibilidade de uma empresa identificar, *a priori*, os fatores que possam influenciar os retornos financeiros futuros gerados por sua atividade operacional.

Segundo Brito e Assaf Neto (2008), por meio das demonstrações contábeis, é possível desenvolver ações que permitem analisar a estrutura e evolução do patrimônio, a liquidez, o endividamento, o retorno do investimento e a lucratividade da empresa. Os autores afirmam que a análise das demonstrações financeiras visa ao estudo do desempenho econômico-financeiro da companhia em determinado

período do passado, para diagnosticar sua posição atual e produzir resultados que sirvam de base para a previsão de tendências futuras.

Os índices financeiros são métricas derivadas das demonstrações contábeis das empresas, os quais podem delinear seu perfil, estratégias competitivas, características econômicas, operacionais e financeiras, bem como decisões sobre investimentos. Tais decisões requerem uma avaliação de mudanças no desempenho ao longo do tempo para um determinado investimento e uma comparação entre todas as empresas dentro de uma única indústria, em um ponto específico no tempo (WHITE; SONDHI; FRIED, 2002).

O nível de retorno gerado por uma empresa pode ser medido por índices financeiros de rentabilidade, tais como *Return on Equity* (ROE), *Return on Assets* (ROA) e *Return on Investment* (ROI). Estes índices são importantes para avaliar a performance de uma companhia e consideradas métricas importantes para a análise de empresas (VAN HORNE; WACHOWICZ, 2008).

O indicador ROE visa apontar o nível de lucro gerado pela empresa em suas atividades operacionais, em relação ao capital investido pelos acionistas. Neste sentido, Martins (2001) advoga que os índices de rentabilidade permitem a avaliação da gestão de recursos próprios e de terceiros, em benefício dos sócios.

Segundo Wernke (2008), a utilização da medida ROA para avaliação de rentabilidade pode identificar a maneira pela qual a margem do lucro aumenta ou se deteriora, a possibilidade de medir a eficiência dos ativos em produzir vendas e a possibilidade de avaliar a gestão do capital de giro. No que se refere à métrica ROI, o autor destaca a possibilidade de compará-lo com a taxa de retorno de outros investimentos, internos ou externos à companhia.

Ademais, Brito e Assaf Neto (2008) afirmam que a situação econômico-financeira futura de uma empresa é um elemento que apresenta significativo grau de incerteza, uma vez que depende de um conjunto amplo de variáveis relacionadas a fatores sistêmicos, como condições econômicas e setoriais, além de fatores específicos da empresa, como sua estrutura e poder de mercado.

Neste ínterim, a aplicação de um modelo teórico para a estimativa da rentabilidade de uma companhia com base em índices financeiros, pode conceder às empresas, subsídios para seu planejamento estratégico e a possibilidade de estimar sua rentabilidade mediante diferentes cenários econômicos, além de permitir a comparação de seu desempenho em relação aos concorrentes.

Altman (1968) desenvolveu um dos primeiros modelos preditivos de falência para empresas, a partir de índices financeiros derivados de demonstrações financeiras e da utilização da técnica estatística de Análise Discriminante para reconhecimento de padrões. Ademais, Ante e Ana (2013) afirmam que os modelos preditivos utilizados para previsão de falência e classificação de risco (*rating*) de empresas de vários setores econômicos costumam empregar técnicas multivariadas.

Os modelos para previsão de resultados financeiros, quando utilizados de forma consistente, têm a capacidade de prever problemas corporativos e conceder subsídios para as decisões estratégicas da companhia. Estas previsões podem ser utilizadas para recomendação de políticas de investimento adequadas e rastreamento de investimentos indesejáveis (ALTMAN, 1993).

De tal modo, a aplicação de técnicas multivariadas como a Análise Discriminante e a Regressão Logística, pode permitir a identificação dos índices financeiros que possuem maior relevância sobre a rentabilidade de empresas do setor automotivo. Portanto, torna-se possível a avaliação dos níveis de retorno de uma empresa a partir de informações oriundas de demonstrações financeiras.

Finalmente, o reconhecimento de padrões de rentabilidade em empresas automotivas possibilita que uma companhia seja classificada em relação a seus concorrentes diretos, no que tange aos resultados gerados por suas atividades operacionais.

1.1 PROBLEMA DE PESQUISA

Em meados do século XX, observou-se o advento de um novo conceito de produção relacionado a um conjunto de inovações organizacionais implantados pela montadora japonesa Toyota, dando origem à filosofia *Lean Manufacturing* ou Produção Enxuta. Este paradigma busca a otimização dos sistemas de produção e aumento da rentabilidade, por meio da eliminação de desperdícios, redução de custos, maior eficiência e melhores resultados financeiros (STEVENSON, 2001).

Em um ambiente de alta competitividade, as empresas automotivas podem buscar estratégias de diferenciação em relação a seus concorrentes. Segundo Porter (1986), existem dois tipos básicos de vantagem competitiva: baixo custo e diferenciação. Tais fatores resultam em três estratégias para alcançar desempenho superior, a saber: liderança em custo, diferenciação e foco.

As montadoras de veículos se caracterizam por estarem voltadas ao alcance de economias de escala, por meio da especialização por plataforma de automóvel e pela flexibilidade permitida pela organização na forma modular. Ademais, no período entre 1994 e 2002, notou-se o aumento de investimentos no setor automotivo brasileiro, bem como a busca por fusões e *joint-ventures*, de modo a obter a modernização de tecnologias e práticas de mercado internacionais (MDIC, 2015).

A análise da rentabilidade de empresas automotivas pode revelar sua eficiência em aproveitar oportunidades de negócios em um ambiente altamente incerto. Neste sentido, os índices financeiros visam capturar os fundamentos do desempenho do negócio de uma forma holística, olhando para os resultados financeiros e sua posição patrimonial.

No entendimento de Jarillo (2003), manter ou aumentar os lucros dos acionistas é necessariamente o propósito estratégico de toda companhia e a capacidade de uma empresa para executar seus compromissos reforça a sua credibilidade junto aos acionistas, clientes, fornecedores e demais *stakeholders*.

À medida que os investidores não são capazes de conhecer antecipadamente possíveis quedas nos resultados operacionais, há possibilidade de se subestimar a situação financeira da empresa, a qual poderá, eventualmente, declarar falência.

Deste modo, apresenta-se o problema de pesquisa a ser investigado: Como avaliar a rentabilidade de uma empresa do setor automotivo?

1.2 OBJETIVOS

Nesta seção serão apresentados o objetivo geral e os objetivos específicos a serem abordados ao longo do presente trabalho.

1.2.1 Objetivo geral

O objetivo desta pesquisa é analisar a rentabilidade de empresas do setor automotivo por meio de técnicas multivariadas, a fim de subsidiar decisões de planejamento estratégico.

Para consecução do objetivo geral serão definidos os objetivos específicos.

1.2.2 Objetivos específicos

a) Coletar índices econômico-financeiros de montadoras de veículos;
b) Separar as empresas em grupos (*clusters*), de acordo com seus índices de rentabilidade, por meio da técnica multivariada de Análise de Agrupamentos;
c) Identificar as variáveis relevantes ou discriminantes para a rentabilidade das empresas, através dos métodos de Regressão Logística e Análise Discriminante de Fisher;
d) Comparar os métodos de Regressão Logística e Análise Discriminante de Fisher para reconhecimento de padrões na rentabilidade das montadoras de veículos;
e) Classificar uma empresa quanto à sua rentabilidade;
f) Parametrizar o modelo selecionado em um simulador desenvolvido na linguagem *Visual Basic for Applications*, no aplicativo *Microsoft Excel*.

1.3 JUSTIFICATIVA

O desenvolvimento de técnicas que apoiem as decisões estratégicas no âmbito da Engenharia de Produção, em especial na área financeira, pode ser visto pelas companhias como um fator de competitividade. A possibilidade de analisar a rentabilidade de empresas pode promover benefícios para acionistas, clientes e investidores, além de estimular a criação de mecanismos para análise da viabilidade econômica de novos projetos sob diferentes cenários.

Ademais, o setor automotivo brasileiro possui singular importância para a economia nacional, uma vez que é responsável por aproximadamente 23% do PIB industrial do país e gera ampla quantidade de empregos diretos e indiretos, em diversas regiões do Brasil (MDIC, 2015).

Sob a ótica social, modelos preditivos de rentabilidade para as organizações deste segmento podem contribuir para maior solidez financeira e, consequentemente, para a manutenção de seu nível de empregabilidade.

Este trabalho procura proporcionar, ainda, o fomento a pesquisas voltadas à análise financeira em Programas de Pós Graduação em Engenharia de Produção, com utilização de métodos estatísticos. Além disto, espera-se suprir uma lacuna

existente na literatura, ao introduzir o conceito de modelos preditivos aplicados à análise de rentabilidade de montadoras de veículos.

Por fim, conforme classificação disponibilizada pela Associação Brasileira de Engenharia de Produção (ABEPRO), o tema desta pesquisa possui aderência à Engenharia de Produção, especificamente, na ramificação de Gestão Financeira da área de Engenharia Econômica.

1.4 LIMITAÇÕES DO TRABALHO

Esta pesquisa se restringe à aplicação de modelos estatísticos para análise financeira de empresas, cuja atividade principal consiste na montagem de veículos. Portanto, fabricantes de autopeças e empresas destinadas a projeto, desenvolvimento e distribuição de automóveis, não estão incluídas nesta pesquisa.

Os modelos aplicados neste trabalho foram calibrados conforme os índices financeiros das empresas automotivas em estudo. As variáveis explicativas dos modelos podem variar significativamente conforme o segmento da economia a ser analisado, haja vista a peculiaridade dos indicadores financeiros para cada setor.

No que se refere à integridade das informações utilizadas, ressalta-se que os índices financeiros necessários ao desenvolvimento dos objetivos deste trabalho foram diretamente extraídos do sistema *MorningStar*®. Ademais, não foi possível obter informações acerca dos padrões contábeis internacionais adotados nas demonstrações financeiras disponibilizadas pela plataforma.

Este trabalho visa desenvolver um mecanismo para análise financeira. O foco desta pesquisa consiste na utilização de modelos preditivos a partir de informações de demonstrações contábeis, de modo que variáveis qualitativas não fazem parte da abrangência deste trabalho. Todavia, espera-se que aspectos como o tempo de maturação das empresas, sua cultura organizacional e sua posição no mercado sejam refletidos por meio dos índices financeiros selecionados.

1.5 ESTRUTURA DO TRABALHO

A fim de alcançar os objetivos delineados na estruturação da pesquisa, o presente trabalho encontra-se segregado em cinco capítulos.

No primeiro capítulo são apresentados os elementos referentes ao planejamento da pesquisa, incluindo introdução, problema de pesquisa, objetivo geral, objetivos específicos, limitações do trabalho e justificativa.

No segundo capítulo abordam-se os principais temas pertinentes ao desenvolvimento do trabalho com base em levantamento bibliográfico, os quais abrangem o panorama do setor automotivo brasileiro, análises de desempenho com base em índices econômico-financeiros e métodos multivariados.

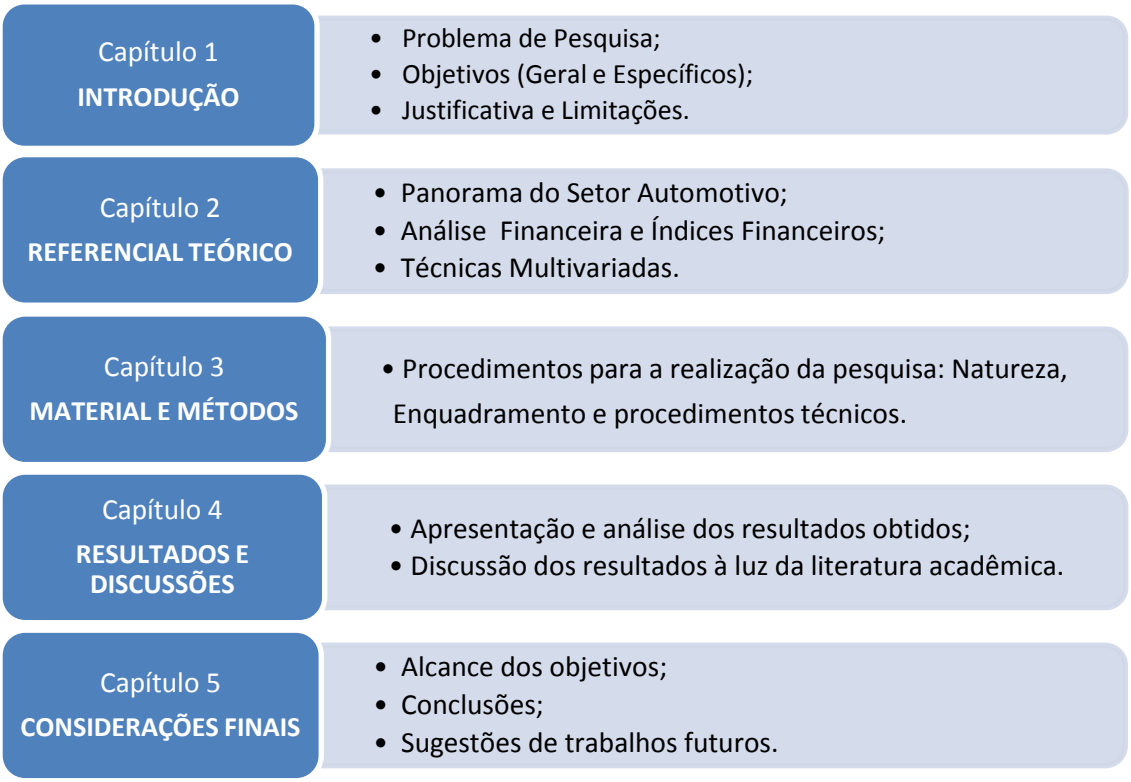
O terceiro capítulo explora os aspectos metodológicos concernentes ao trabalho, tais como a natureza e o enquadramento da pesquisa. Ademais, são descritos os procedimentos utilizados para a coleta de dados e o detalhamento sobre a aplicação de técnicas multivariadas ao problema estudado.

No quarto capítulo são apresentadas a análise, a interpretação e a contextualização dos resultados alcançados no trabalho.

O quinto capítulo, por fim, destina-se às considerações finais relativas às análises apresentadas no capítulo anterior.

A FIGURA 1 ilustra a estrutura do trabalho, bem como as etapas a serem executadas em cada capítulo.

FIGURA 1 - ESTRUTURA DO TRABALHO



FONTE: O autor (2016).

2 REFERENCIAL TEÓRICO

2.1 O PANORAMA DO SETOR AUTOMOTIVO

O cenário de desaceleração econômica nos países ocidentais, na década de 1950, demandou a flexibilização dos sistemas de produção para melhor atendimento dos clientes (SLACK; CHAMBERS; JOHNSTON, 2002). De tal modo, a montadora japonesa *Toyota* propôs uma nova filosofia, denominada *Lean Manufacturing*, buscando alternativas para produção eficiente em um mercado de pequenas dimensões, baixa produtividade e escassez de recursos.

Na visão de Stevenson (2001), as empresas que adotaram o sistema *Lean Manufacturing* apresentaram, em geral, vantagens competitivas sobre as companhias que empregavam abordagens tradicionais, não somente em países desenvolvidos, mas também em nações emergentes.

Os fabricantes de automóveis costumam dispor de estruturas globais, em decorrência do alto padrão tecnológico do setor e volume do mercado, que possui elevados custos de desenvolvimento de novos produtos, *setup*, investimentos em máquinas, equipamentos e infraestrutura produtiva (MDIC, 2015).

O setor automotivo brasileiro, ao longo dos anos, busca atrair investimentos internacionais de montadoras de veículos. Observa-se na FIGURA 2, a posição de destaque do Brasil entre os maiores mercados consumidores.

FIGURA 2 - RANKING DE PRODUÇÃO DE VEÍCULOS (2008 - 2014)

	Países	2014		2013		2012		2011		2010		2009		2008
1º	China	21.004.688	1º	19.311.225	1º	16.366.208	1º	15.237.749	1º	14.834.259	2º	9.848.074	2º	6.492.553
2º	EUA	16.517.204	2º	15.597.227	2º	14.495.293	2º	12.778.868	2º	11.589.672	1º	10.418.730	1º	13.221.559
3º	Japão	5.495.939	3º	5.320.994	3º	5.320.391	3º	4.170.277	3º	4.919.718	3º	4.577.288	3º	5.032.330
4º	Brasil	3.328.958	4º	3.575.947	4º	3.634.627	4º	3.425.495	4º	3.328.254	5º	3.011.285	6º	2.670.852
5º	Alemanha	3.261.376	5º	3.161.669	5º	3.298.413	5º	3.403.514	5º	3.109.659	4º	3.982.467	4º	3.318.311
6º	Índia	2.889.595	6º	2.885.509	6º	3.097.285	6º	2.802.485	7º	2.640.018	9º	1.967.472	10º	1.675.021
7º	Grã-Bretanha	2.798.121	8º	2.535.810	8º	2.284.250	9º	2.201.406	8º	2.253.761	8º	2.181.387	8º	2.421.256
8º	Rússia	2.485.924	7º	2.777.601	7º	2.935.233	7º	2.653.676	10º	1.910.765	10º	1.465.925	5º	2.925.401
9º	França	2.167.951	9º	2.157.880	9º	2.282.816	8º	2.633.487	6º	2.669.285	6º	2.642.657	7º	2.510.555
10º	Canadá	1.853.047	10º	1.747.088	10º	1.678.039	11º	1.587.512	11º	1.558.572	11º	1.459.735	11º	1.637.839

FONTE: FENABRAVE (2015).

Não obstante, a FIGURA 3 denota a desaceleração na produção de veículos no Brasil ao longo dos últimos anos.

FIGURA 3 - PRODUÇÃO DE VEÍCULOS NO BRASIL: VARIAÇÃO ANUAL(%)

	Total	Automóveis e Comerciais Leves	Caminhões	Ônibus	Motos	Implementos
2005	11,3	10,5	-4,3	-13,4	16,4	-11,4
2006	15,7	12,2	-3,9	27,6	23,5	0,9
2007	24,2	22,8	22,0	1,2	26,8	17,5
2008	14,2	14,1	24,9	18,9	12,7	34,3
2009	-0,1	23,6	-11,4	-14,3	-16,4	-10,2
2010	12,4	10,6	44,4	25,3	12,1	46,2
2011	5,0	2,9	9,7	21,9	7,6	13,0
2012	2,3	6,1	-20,2	-15,1	-15,6	3,8
2013	-2,3	-1,6	13,0	19,6	-8,5	17,9
2014	-6,9	-11,3	-12,7	-5,7	-18,2	-15,6
2015	-15,8	-15,7	-25,0	-13,0	-14,0	-27,0

FONTE: FENABRAVE (2015).

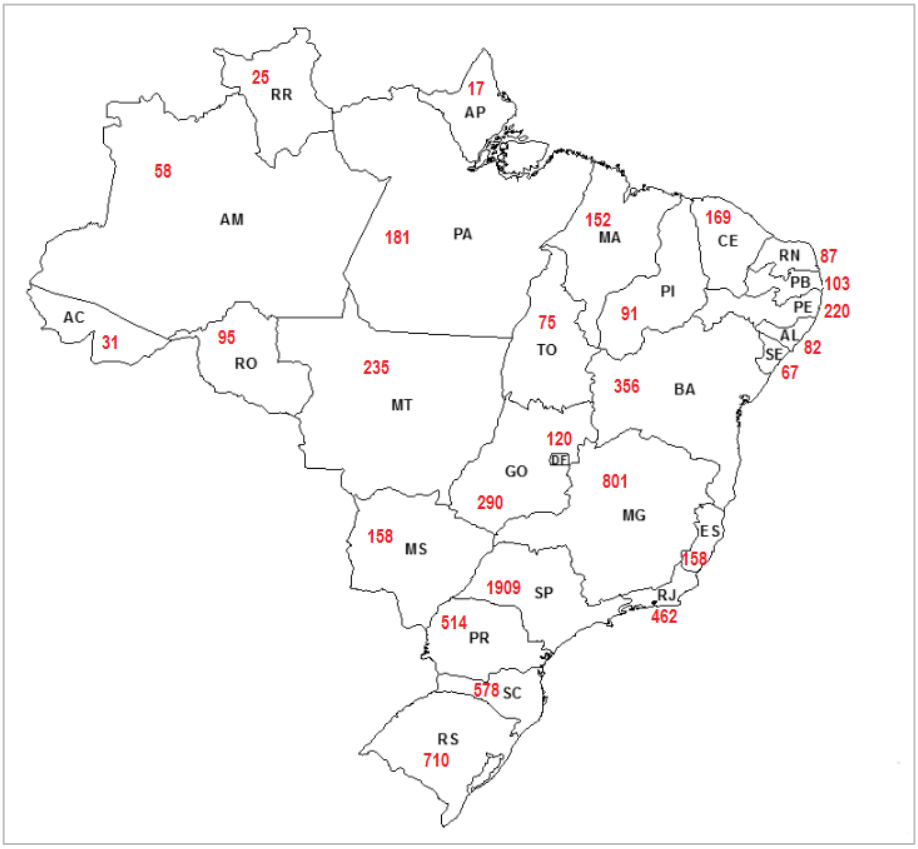
Em janeiro de 2013, um novo conjunto de medidas anunciadas pelo governo visou alavancar a produção de peças e direcionar uma parcela das receitas para as áreas de Pesquisa e Desenvolvimento (P&D). Na ocasião, os fabricantes que investiram anualmente um percentual de seu faturamento bruto em atividades de pesquisa e desenvolvimento em Tecnologia da Informação foram beneficiados.

A rede de concessionárias, por sua vez, desenvolve um importante papel para o setor automotivo, pois os canais de distribuição formam um conjunto de organizações interdependentes envolvidas no processo de tornar produtos e serviços disponíveis para o consumo (STERN; EL-ANSARY; COUGHAN, 1996).

Neste sentido, a seleção e o gerenciamento dos canais de distribuição englobam a construção de uma série de mecanismos e de uma rede por meio da qual a empresa alcança seu público alvo mercado, mantém-se em contato com seus clientes e realiza uma série de atividades fundamentais, que vão desde a geração de demanda até a entrega física dos produtos.

A FIGURA 4 apresenta a distribuição das concessionárias em todo o território brasileiro.

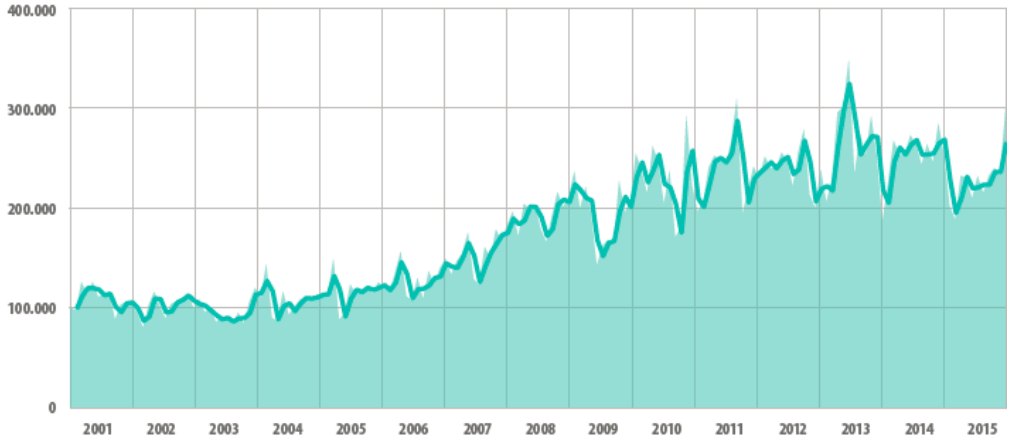
FIGURA 4 - CONCESSIONÁRIAS ASSOCIADAS À FENABRAVE



FONTE: FENABRAVE (2015).

O GRÁFICO 1 mostra dados históricos referentes à quantidade de emplacamentos de automóveis.

GRÁFICO 1 - EMPLACAMENTOS DE AUTOMÓVEIS NO BRASIL POR ANO



FONTE: FENABRAVE (2015).

Na TABELA 1 será apresentada a distribuição das principais montadoras atualmente instaladas no Brasil, em termos de seu *market share*. Nota-se a forte participação de montadoras tradicionais, como Fiat e General Motors.

TABELA 1 - *MARKET SHARE*: AUTOMÓVEIS E COMERCIAIS LEVES

MONTADORA	*MARKET SHARE*
FIAT	20,40%
GM	19,20%
VOLKSWAGEN	16,20%
FORD	10,20%
HYUNDAI	7,20%
RENAULT	6,30%
TOYOTA	5,10%

FONTE: FENABRAVE (2015).

Por fim, a indústria brasileira de automóveis pode ser resumida da seguinte forma (MDIC, 2015):

- 31 fabricantes de veículos e de máquinas agrícolas e rodoviárias;
- 500 fabricantes de autopeças;
- 5.533 concessionárias;
- 64 unidades industriais em 10 estados e 52 municípios;
- Capacidade produtiva instalada de 4,5 milhões de unidades de veículos;
- Capacidade produtiva instalada de 100 mil unidades de máquinas agrícolas e rodoviárias;
- Faturamento (incluindo autopeças) de US$ 110,9 bilhões em 2014;
- Investimentos na ordem de US$ 68 bilhões no período entre 1994 e 2012;
- Produção acumulada 71,2 milhões de unidades de veículos, e 2,5 milhões de máquinas agrícolas e rodoviárias, no período entre 1960 e 2014;
- Exportações de US$ 18,5 bilhões em 2012;
- Importações de US$ 30,2 bilhões (incluindo autopeças) em 2014;
- 1,5 milhão de empregos diretos e indiretos;
- Participação de 23% no PIB Industrial e 5% no PIB Total;
- Ranking mundial em 2014: 4º maior mercado consumidor.

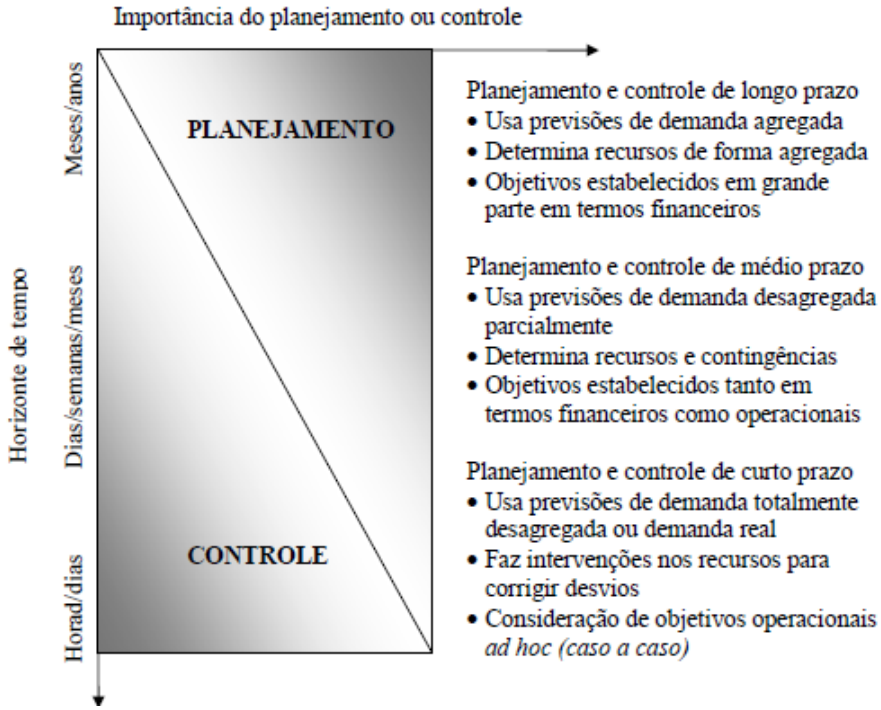
Neste contexto, a administração financeira, mediante análise de demonstrações contábeis e índices financeiros, pode ser um importante instrumento para acompanhamento dos resultados destas empresas e mapeamento dos fatores-chave que influenciam seus resultados.

2.2 O PAPEL DA ADMINISTRAÇÃO FINANCEIRA NO PLANEJAMENTO ESTRATÉGICO

O objetivo da Administração Financeira consiste na maximização do capital dos proprietários, de modo que a tarefa mais importante de um administrador financeiro é criar valor a partir das atividades de orçamento de capital (ROSS; WESTERFIELD; JAFFE, 2002).

Slack, Chambers e Johnston (2002) consideram o Plano Mestre de Produção como a fase mais importante do Planejamento e Controle da Produção (PCP), pois contém a definição acerca da quantidade e do momento em que os produtos finais devem ser produzidos. Os autores afirmam que este programa direciona a operação e a utilização de recursos, além de determinar o aprovisionamento de materiais e capital, conforme apresentado na FIGURA 5.

FIGURA 5 - PLANEJAMENTO E CONTROLE DE PRODUÇÃO



FONTE: SLACK; CHAMBERS; JOHNSTON (2002).

Visto que os recursos à disposição da organização são limitados, o Planejamento Estratégico e o Plano Mestre de Produção são delineados levando-se em consideração a alocação e gerenciamento destes recursos.

Ressalta-se, ademais, a importância do engajamento entre as diversas áreas da organização com a Produção, dentre as quais se inclui a área financeira. Neste contexto, metodologias voltadas ao nível estratégico das organizações, tais como o *Lean Manufacturing* e *Balanced Score Card (BSC)* apregoam a sinergia gerada pela interação conjunta dos departamentos, de forma a estabelecer processos de melhoria contínua e agregar valor para a organização.

2.3 ANÁLISE DA ESTRUTURA PATRIMONIAL DE UMA ORGANIZAÇÃO

A situação econômico-financeira de uma empresa pode ser avaliada por suas demonstrações contábeis, evidenciando aspectos tais como a disponibilidade de recursos em caixa para honrar compromissos de curto prazo, o grau de utilização de recursos de terceiros para financiamento de suas operações e o retorno sobre o capital investido pelos acionistas (BREALEY; MYERS; MARCUS, 2001).

Os relatórios financeiros também devem fornecer informações sobre a maneira pela qual a administração cumpriu sua função junto aos acionistas, no que se refere ao uso dos recursos da empresa. Assim, a administração não é responsável somente pela guarda e custódia de recursos, mas também por seu uso eficiente e rentável (WHITE; SONDHI; FRIED, 2002).

A análise tradicional das demonstrações financeiras considera os índices de liquidez, endividamento, atividade e rentabilidade como parâmetros para avaliação de empresas. A lei 6.404/76 apresenta as demonstrações que devem ser publicadas por sociedades anônimas, as quais sejam: Balanço Patrimonial (BP), Demonstração do Resultado do Exercício (DRE), Demonstração das Mutações do Patrimônio Líquido (DMPL) e Demonstração das Origens e Aplicações de Recursos (DOAR).

2.3.1 Balanço patrimonial

Trata-se de uma ferramenta acerca da posição financeira de uma empresa em um determinado instante de tempo, compreendendo bens, direitos e obrigações de curto e longo prazo.

O Balanço Patrimonial evidencia, resumidamente, o patrimônio da entidade de modo a facilitar o conhecimento e a análise da situação financeira da companhia (RIBEIRO, 2009). Esta demonstração financeira é composta por ativos, passivos e pelo patrimônio líquido.

Os ativos representam os bens e direitos da companhia, segregados em ativos circulantes e não circulantes, de forma a refletir seu grau de liquidez. Estas contas são segregadas em:

a) Ativo Circulante: formado por contas que estão em constante movimento, e bens ou direitos que serão convertidos em dinheiro em até 12 meses contados a partir da data de referência do balanço. Esta categoria de ativos inclui valores em caixa (disponibilidades), contas bancárias, recebíveis de curto prazo e estoques;
b) Ativo Realizável a Longo Prazo: engloba os bens e os direitos a serem convertidos em dinheiro após 12 meses contados a partir da data de referência do balanço. Inclui contas a receber referente às vendas, empréstimos a sócios, entre outros;
c) Investimentos: aplicações financeiras que geram rendimentos, mas não fazem parte das atividades operacionais da empresa;
d) Ativo Imobilizado: constituído por bens de natureza permanente, os quais são utilizados para a manutenção das atividades operacionais da empresa, tais como veículos, máquinas e imóveis;
e) Ativo Diferido: formado por contas que afetam resultados de exercícios futuros.

Os passivos compreendem obrigações de curto e longo prazo, tais como financiamentos, empréstimos e compromissos junto a fornecedores. Estas contas são divididas em três categorias, a saber:

a) Passivo Circulante: constituído por obrigações exigíveis, que deverão ser liquidadas no prazo máximo de 12 meses contados a partir da data de referência do balanço, tais como fornecedores a pagar, salários a pagar, empréstimos e financiamentos;

b) Passivo Exigível a Longo Prazo: formado pelas obrigações que serão liquidadas com prazo superior a 12 meses contados a partir da data de referência do balanço;
c) Resultado de Exercícios Futuros: corresponde às receitas que são recebidas antecipadamente.

O Patrimônio Líquido é formado por recursos dos acionistas investidos na empresa, além dos lucros ou prejuízos acumulados ao longo do exercício contábil.

O total de passivos, acrescido do Patrimônio Líquido, compreende as dívidas junto aos credores externos, fornecedores ou bancos e as verbas remanescentes que são devidas aos acionistas, incluindo os lucros acumulados reinvestidos no negócio (ASSAF NETO, 2007).

2.3.2 Demonstração do resultado do exercício

A Demonstração do Resultado do Exercício (DRE) evidencia a formação do resultado líquido do exercício, através do confronto de receitas e despesas. A DRE oferece uma síntese dos resultados operacionais e não operacionais de uma empresa em certo período.

As receitas auferidas pela empresa são divididas em:

a) Receita bruta de vendas: corresponde às receitas decorrentes da venda de produtos ou serviços comercializados pela empresa;
b) Receita líquida de vendas: Receita bruta de vendas, deduzida de tributos incidentes sobre o faturamento, devoluções e descontos concedidos a clientes;
c) Receitas financeiras: receitas derivadas de transações financeiras, tais como os juros de mora recebidos no período;
d) Receitas operacionais: corresponde às receitas relacionadas com a atividade principal da empresa, ou seja, os valores pelos quais a empresa procura ressarcir os custos incorridos na produção dos bens a serem comercializados (Custo das Mercadorias Vendidas);
e) Receitas não operacionais: corresponde a outras receitas, as quais não estão relacionadas à atividade principal da empresa.

As despesas, por sua vez, são segregadas em:

a) Despesas operacionais: gastos gerais com a administração da empresa, tais como aluguéis, materiais de escritório e distribuição de produtos;
b) Despesas financeiras: englobam a remuneração paga às fontes de financiamento da empresa (capitais de terceiros), tais como juros, comissões, variação cambial e despesas bancárias;
c) Despesas não operacionais: não estão diretamente relacionadas com as atividades operacionais da empresa, incluindo prejuízo com vendas de imobilizado e investimentos.

O exemplo da TABELA 2 ilustra a composição de uma DRE:

TABELA 2 - EXEMPLO DE DEMONSTRAÇÃO DO RESULTADO DO EXERCÍCIO

CONTA CONTÁBIL	VALOR (R$)
RECEITA BRUTA DE VENDAS	1.000
(-) DEDUÇÕES DE VENDAS	100
(=) RECEITA LÍQUIDA DE VENDAS	**900**
(-) CUSTO DAS MERCADORIAS VENDIDAS	200
(=) RECEITA OPERACIONAL	**700**
(-) DESPESAS OPERACIONAIS	50
(=) LUCRO OPERACIONAL	**650**
(+) RECEITAS FINANCEIRAS	100
(-) DESPESAS FINANCEIRAS	50
(+) RECEITAS NÃO OPERACIONAIS	40
(-) DESPESAS NÃO OPERACIONAIS	20
(=) LUCRO ANTES DE IRPJ E CSLL	**720**
(-) PROVISÃO PARA IRPJ E CSLL	288
(-) PARTICIPAÇÕES E CONTRIBUIÇÕES	200
(=) LUCRO LÍQUIDO DO EXERCÍCIO	**232**

FONTE: Adaptado de MARION (2007).

Na visão de Matarazzo (2010), a DRE é o resumo do movimento de certas receitas e despesas no balanço entre duas datas. Segundo o autor, a principal característica do relatório é demonstrar o resultado apresentado pela atividade operacional da organização em determinado período.

2.4 ÍNDICES ECONÔMICO-FINANCEIROS

De modo geral, os índices expressam uma relação matemática entre quantidades, a qual pode revelar condições e tendências que muitas vezes não podem ser observadas por intermédio de seus componentes individuais.

Segundo Gitman e Madura (2003), a análise de índices financeiros visa examinar e monitorar o desempenho das empresas e possui como partes interessadas: os acionistas, que estudam os níveis de risco e retorno; os credores, que avaliam a liquidez de curto prazo e a capacidade de pagamento; e a administração, que tem o objetivo de produzir índices financeiros que sejam favoráveis aos outros usuários e monitorar o desempenho da empresa.

A análise financeira, por meio de informações coletadas a partir de demonstrações contábeis, pode ser utilizada por investidores, clientes e outros *stakeholders* para a análise da saúde de uma empresa, bem como o mapeamento das variáveis que afetam seus resultados de forma significativa.

No contexto do planejamento estratégico, os dados das demonstrações contábeis são utilizados na análise do desempenho global da empresa e na avaliação de sua situação financeira atual, tendo como objetivos verificar se a conjuntura da própria organização se encontra dentro dos parâmetros do setor, compreender as políticas seguidas por um concorrente ou verificar a saúde financeira de um cliente (BREALEY; MYERS; MARCUS, 2001).

Deste modo, Porter (1986) define a estratégia corporativa como o plano geral que direciona toda a organização, tais como decisões relativas à estrutura de capital e a priorização de projetos de acordo com os recursos disponíveis. Segundo o autor, as decisões estratégicas acompanham a estratégia geral da empresa, a partir de medidas com foco em suas operações.

Os índices financeiros podem ser utilizados para examinar o desempenho atual de uma companhia, em comparação com períodos anteriores, gerar subsídios aos gestores acerca de dificuldades que necessitam ser sanadas e embasar

decisões estratégicas relativas ao processo produtivo da empresa. Destarte, há possibilidade de identificar problemas potenciais e delinear ações preventivas para mitigação destes riscos.

Todas as variáveis comuns utilizadas nas fórmulas seguintes encontram-se definidas nas seções 2.3.1 e 2.3.2.

2.4.1 Índices de liquidez

Na visão de Van Horne e Wachowicz (2008), os índices de liquidez são utilizados para medir a capacidade da empresa em cumprir suas obrigações. A partir destas medidas, é possível se obter uma visão sobre a atual condição de uma companhia e sua capacidade de permanecer solvente em caso de adversidades.

Os índices de liquidez constituem uma apreciação a respeito da capacidade da empresa em saldar seus compromissos (MARION, 2007). Por meio da análise destes indicadores, é possível analisar a gestão de fluxos de caixa da companhia, uma vez que a mesma deve possuir recursos financeiros para liquidar seus compromissos na data de seu vencimento.

2.4.1.1 Liquidez geral

Segundo Assaf Neto (2007), este quociente é utilizado para detectar a saúde financeira da empresa, indicando a quantidade de ativos circulantes e de ativos realizáveis a longo prazo para cada unidade monetária de dívida total, sendo utilizada como uma medida de segurança financeira e de capacidade para honrar todos os compromissos assumidos. O índice é expresso por:

$$Liquidez\ Geral = \frac{Ativo\ Circulante\ + Ativo\ Realizável\ a\ Longo\ Prazo}{Passivo\ Circulante\ + Passivo\ Exigível\ a\ Longo\ Prazo} \quad (2.1)$$

2.4.1.2 Liquidez corrente

Este índice reflete a capacidade da empresa para cobrir suas atuais responsabilidades com seus ativos correntes. Segundo Brealey, Myers e Marcus

(2001), reduções rápidas neste índice podem ser reflexo de alguma dificuldade. O índice é calculado por:

$$Liquidez\ Corrente = \frac{Ativo\ Circulante}{Passivo\ Circulante} \quad (2.2)$$

O coeficiente relaciona o montante de recursos financeiros, bens e direitos realizáveis a curto prazo com as obrigações a serem pagas dentro do mesmo período. Quanto maior for a liquidez corrente, mais alta será a capacidade da empresa em financiar necessidades de capital de giro (ASSAF NETO, 2007).

2.4.1.3 Liquidez seca

Este índice demonstra a porcentagem das dívidas de curto prazo em condições de serem saldadas mediante a utilização dos itens monetários de maior liquidez. Essencialmente, a liquidez seca determina a capacidade de curto prazo de pagamento da empresa mediante a utilização das contas de disponibilidades e valores a receber (ASSAF NETO, 2007).

Havendo necessidade em liquidar alguma obrigação tempestivamente, a empresa não deve vender seu estoque de produtos acabados acima do valor de seu custo de produção. Assim, os estoques podem ser excluídos da comparação entre o Ativo Circulante e o Passivo Circulante (BREALEY; MYERS; MARCUS, 2001).

O coeficiente revela a capacidade da empresa em fazer frente ao Passivo Circulante com seus ativos mais líquidos, ou seja:

$$Liquidez\ Seca = \frac{Ativo\ Circulante - Estoques}{Passivo\ Circulante} \quad (2.3)$$

Considera-se, portanto, que os ativos prontamente disponíveis são rapidamente conversíveis em caixa sem grandes perdas de valor. Portanto, este indicador mede a capacidade da empresa de saldar suas dívidas imediatas e de curto prazo, sem abrir mão de seus estoques.

2.4.1.4 Liquidez imediata

Os ativos mais líquidos de uma empresa são suas disponibilidades, ou seja, os recursos em caixa e títulos (públicos ou privados) negociáveis no mercado secundário (BREALEY; MYERS; MARCUS, 2001).

O índice de liquidez imediata mostra a proporção entre os ativos imediatamente líquidos e passivos de curto prazo. Sua expressão é dada por:

$$Liquidez\ Imediata = \frac{Disponibilidades}{Passivo\ Circulante} \tag{2.4}$$

2.4.2 Índices de endividamento

Os índices de endividamento visam analisar o nível de recursos dos proprietários e de terceiros que são consumidos nas atividades da empresa. Parte-se do princípio de que na medida em que a empresa possui mais dívidas, maior é a possibilidade de não cumprimento de suas obrigações contratuais.

No entendimento de Brealey, Myers e Marcus (2001), estes índices fornecem informações relevantes sobre a saúde financeira de uma empresa, visto que a mesma deve dispor de recursos suficientes para cobrir o fluxo de juros e principal a ser pago a seus credores para captar recursos para implementação de projetos ou para aquisição de bens de capital.

Neste sentido, a análise destes índices pode medir o nível de endividamento de uma empresa, bem como oferecer o conhecimento a respeito de sua capacidade em reembolsar o saldo devedor de suas dívidas, pagar juros sobre seus empréstimos, e satisfazer as suas outras obrigações financeiras com recursos próprios ou de terceiros.

2.4.2.1 Endividamento geral

O índice de endividamento geral, também conhecido por índice de endividamento total, apresenta a relação entre o total de direitos e obrigações de uma empresa (ASSAF NETO, 2007).

A variável mensura o percentual de recursos gerados por dívidas junto a fontes de financiamento, sendo definida por:

$$Endividamento\ Geral = \frac{Passivo\ Total}{Ativo\ Total} \quad (2.5)$$

2.4.2.2 Endividamento de longo prazo

O grau de endividamento de uma companhia, geralmente, é medido pela relação entre a dívida de longo prazo e o total de bens e direitos realizáveis a longo prazo (BREALEY; MYERS; MARCUS, 2001). Este índice é calculado por:

$$Endividamento\ de\ Longo\ Prazo = \frac{Passivo\ Exigível\ a\ Longo\ Prazo}{Ativo\ Total} \quad (2.6)$$

2.4.2.3 Participação de capitais de terceiros

Segundo Assaf Neto (2007), a Participação de Capitais de Terceiros (PCT) relaciona a soma entre o Passivo Circulante e o Passivo Exigível a Longo Prazo, com os fundos totais gerados por recursos próprios e por fontes de financiamento externas. Esta relação é expressa por:

$$PCT = \frac{Passivo\ Circulante + Exigível\ a\ Longo\ Prazo}{Ativo\ Total} \quad (2.7)$$

Ao analisar este coeficiente, há interesse em analisar a empresa sob a ótica de seu risco de insolvência, indicando a porcentagem dos ativos financiados com recursos de terceiros (MATARAZZO, 2010).

2.4.2.4 *Debt to equity* (D/E)

O índice *Debt to Equity* (D/E) revela a proporção relativa de dívida e capital próprio que uma empresa emprega. Este índice pode fornecer a seus *stakeholders*, em especial credores e fornecedores, informações sobre o volume de dívidas da organização. Portanto, tem-se que:

$$D/E = \frac{Passivo\ Circulante + Exigível\ a\ Longo\ Prazo}{Patrimônio\ Líquido} \quad (2.8)$$

Este coeficiente reflete a proteção dos credores contra insolvência e o apetite da companhia para a obtenção de novos financiamentos, tendo em vista o aproveitamento de oportunidades de investimento entendidas como potencialmente atraentes.

Empresas com fluxos de caixa bastante estáveis, geralmente possuem maior proporção da dívida em capital que outras entidades, cujos fluxos de caixa sejam mais voláteis. A comparação entre companhias de características similares fornece uma indicação geral acerca de sua credibilidade e risco financeiro inerente às suas atividades (VAN HORNE; WACHOWICZ, 2008).

2.4.2.5 Composição do endividamento

Após a avaliação do grau de endividamento da empresa, o índice de Composição do Endividamento (CE) é calculado com a finalidade de medir a composição de suas dívidas. Na medida em que o índice assume valores menores, maior será a proporção de dívidas de longo prazo, concedendo à empresa maior período de tempo para gerar recursos que saldarão seus compromissos. Assim, o índice pode ser expresso por:

$$CE = \frac{Passivo\ Circulante}{Passivo\ Circulante + Exigível\ a\ Longo\ Prazo} \quad (2.9)$$

Caso o índice mostre significativa concentração no Passivo Circulante (obrigações de curto prazo), a empresa poderá apresentar dificuldades em momentos de reversão de mercado, o que não aconteceria caso as dividas estivessem concentradas no longo prazo (MARION, 2007).

2.4.3 Índices de atividade

Também conhecidos como índices de eficiência ou índices de volume de negócios visam medir a eficácia com que a empresa está usando seus ativos.

Alguns aspectos da análise de atividades estão intimamente relacionados à avaliação de liquidez. Em linhas gerais, estes itens abordam a eficácia com que a empresa gerencia, em especial, seus recebíveis e estoques (VAN HORNE; WACHOWICZ, 2008).

2.4.3.1 Giro do ativo total

Segundo Ribeiro (2009), esta medida avalia a proporção existente entre o volume de vendas e os ativos totais da empresa, ou seja, o quanto a empresa vendeu para cada unidade monetária de investimento total.

De tal modo, este índice verifica se o volume das vendas realizadas no período foi adequado em relação ao capital total investido na empresa. Portanto, esta relação é expressa por:

$$Giro\ do\ Ativo\ Total = \frac{Receitas\ Operacionais}{Ativo\ Total} \tag{2.10}$$

Caso o índice seja elevado, em comparação com outras empresas do mesmo setor, pode haver indícios de que a empresa está trabalhando perto de sua capacidade produtiva. Neste caso, pode ser complexo gerar mais negócios sem investimentos adicionais (BREALEY; MYERS; MARCUS, 2001).

2.4.3.2 Giro do ativo imobilizado

Este índice mede a relação entre as vendas líquidas anuais e o total de ativos investidos em imóveis, instalações e equipamentos, após a dedução da respectiva depreciação acumulada. O coeficiente é expresso por:

$$Giro\ do\ Ativo\ Imobilizado = \frac{Receitas\ Operacionais}{Ativo\ Imobilizado} \tag{2.11}$$

Segundo Van Horne e Wachowicz (2008), valores mais elevados são preferíveis, indicando que a empresa possui menos recursos comprometidos em ativos imobilizados para cada unidade monetária da receita operacional gerada por

suas vendas. Índices mais baixos, por sua vez, podem revelar baixo investimento em bens de capital e outros ativos de baixa liquidez.

2.4.3.3 Giro de estoque

O giro de estoque pode ser utilizado para avaliar a eficiência com que a sua empresa utiliza seus ativos para gerar vendas. A empresa torna-se mais eficaz na venda de seus estoques à medida que o estoque médio torna-se menor (MARION, 2007).

No entendimento de Silva (2008), o giro de estoque permite determinar o número de dias (em média) que os produtos permanecem armazenados antes de serem vendidos, sendo expresso por:

$$Giro\ de\ Estoque = \frac{Custo\ das\ Mercadorias\ Vendidas}{Estoques} \tag{2.12}$$

Para que as empresas controlem o ritmo em que seus estoques giram, compara-se o valor de estoques com o Custo das Mercadorias Vendidas, pois as demonstrações financeiras não apresentam o valor de venda dos produtos acabados, mas o custo dos estoques (BREALEY; MYERS; MARCUS, 2001).

2.4.3.4 Prazo médio de recebimento de vendas

O Prazo Médio de Recebimento de Vendas (PMRV) mede a velocidade na qual os clientes quitam seus compromissos, expressando os valores a receber em termos das vendas realizadas (BREALEY; MYERS; MARCUS, 2001). Este índice é expresso por:

$$PMRV = 365 * \frac{Contas\ a\ Receber}{Receitas\ Operacionais} \tag{2.13}$$

Em geral, na medida em que o volume dos valores a receber se torna menor, há indícios de que os clientes efetuam pagamentos à empresa de forma

célere. Por outro lado, uma proporção demasiadamente baixa pode indicar elevado rigor das políticas de crédito da companhia.

2.4.3.5 Prazo médio de pagamento de compras

De acordo com Padoveze (2008), o Prazo Médio de Pagamento de Compras (PMPC) tem a finalidade de medir o prazo médio no qual a empresa é capaz de cumprir seus compromissos junto a fornecedores de materiais e serviços. Esta relação é expressa por:

$$PMPC = \ 365 * \frac{Fornecedores\ a\ Pagar}{Custo\ das\ Mercadorias\ Vendidas} \qquad (2.14)$$

Ademais, este coeficiente mensura quantos dias (em média) a organização leva para pagar compras efetuadas, indicando as condições de crédito obtidas pela empresa junto aos fornecedores (MARION, 2007).

2.4.3.6 Imobilização do patrimônio líquido

Na medida em que a empresa investe em instalações e demais ativos permanentes, menor será seu volume de ativos circulantes e, consequentemente, haverá maior dependência de fontes de financiamento externas (MATARAZZO, 2010).

O grau de imobilização sobre o patrimônio líquido mensura a forma de aplicação dos recursos próprios em ativos imobilizados, sendo expresso por:

$$Imobilização\ do\ Patrimônio\ Líquido\ = \frac{Ativo\ Imobilizado}{Patrimônio\ Líquido} \qquad (2.15)$$

Portanto, este índice demonstra o quanto foi investido na expansão de instalações, para cada unidade monetária de capital investida pelos sócios.

2.4.4 Índices de rentabilidade

A rentabilidade de uma companhia pode ser medida em função dos investimentos e das fontes de financiamento de seus recursos. A administração adequada dos ativos proporciona maior retorno para a empresa (MARION, 2007).

Os índices de rentabilidade revelam a capacidade da geração de fluxos de caixa em relação ao valor investido pelos acionistas. Destarte, estes coeficientes orientam os investidores na avaliação da capacidade de uma empresa para gerar lucro, em comparação com as suas despesas e outros custos relevantes incorridos durante um período específico.

Assaf Neto (2007) afirma que os índices de rentabilidade têm por objetivo avaliar os resultados auferidos por uma empresa em relação a determinados parâmetros que melhor revelam suas dimensões. O autor pondera, ainda, que o embasamento adotado para comparar resultados empresariais costuma ser baseado na análise de ativos, patrimônio líquido e receitas de vendas.

2.4.4.1 Margem operacional

Brealey, Myers e Marcus (2001) definem este índice como a proporção entre as receitas oriundas de vendas e o resultado operacional da empresa. O índice é calculado por:

$$Margem\ Operacional = \frac{Lucro\ Operacional}{Receita\ Líquida\ de\ Vendas} \tag{2.16}$$

O índice mostra o resultado da organização em relação às vendas, após a dedução dos custos de produção dos bens, medindo a eficiência das operações da empresa e os preços de seus produtos (VAN HORNE; WACHOWICZ, 2008).

2.4.4.2 *Return on Assets* (ROA)

Em uma indústria competitiva, as empresas podem obter ganhos suficientes para cobrir apenas seu custo de capital, e o alto retorno sobre os ativos pode ser

visto como um indício de que a empresa está se beneficiando de uma posição de liderança (BREALEY; MYERS; MARCUS, 2001).

Desta forma, o *Return on Assets* mede a eficiência no uso dos ativos por parte da empresa, sendo expresso por:

$$ROA = \frac{Lucro\ Líquido}{Ativo\ Total} \tag{2.17}$$

Segundo Matarazzo (2010), este coeficiente mede o quanto a empresa obtém de lucro para cada unidade monetária de investimento total, indicando o potencial de geração de lucro por parte da empresa.

2.4.4.3 *Return on Investment* (ROI)

O *Return on Investment* (ROI) é uma alternativa à utilização do ROA, a fim de avaliar o retorno produzido por recursos de acionistas e credores aplicados nos negócios. O capital investido é composto pelos recursos onerosos (dívidas da empresa que produzem juros) captados junto a credores, e os recursos próprios aplicados por seus acionistas, cujos valores são registrados em contas do patrimônio líquido (ASSAF NETO, 2007). Esse índice pode ser expresso por:

$$ROI = \frac{Lucro\ antes\ de\ Imposto\ de\ Renda}{Passivo\ Oneroso + Patrimônio\ Líquido} \tag{2.18}$$

Este coeficiente mensura o resultado da empresa ao remunerar seus investimentos totais, relacionando o lucro com o valor dos investimentos realizados. Desta forma, é possível expressar o quanto a empresa gera de lucro para cada unidade monetária de investimento.

De acordo com Wernke (2008), o interesse por este indicador deve-se ao fato de combinar fatores de lucratividade como receitas, custos e investimentos, sendo possível gerar comparações com a taxa de retorno de outros investimentos, internos ou externos à companhia.

2.4.4.4 *Return on Equity* (ROE)

O *Return on Equity* (ROE) visa medir, de forma eficiente, se uma companhia possui capacidade de investir o capital aportado pelos acionistas, gerar lucros e expandir os negócios. Ao contrário de outros índices de retorno sobre o investimento, o ROE está intimamente ligado à visão dos acionistas.

De acordo com Assaf Neto (2007), este coeficiente mensura os retornos obtidos pela empresa a partir dos recursos aplicados por seus acionistas, ou seja, reflete o quanto os proprietários recebem de retorno para cada unidade monetária de recursos próprios investida na empresa.

O índice não deve ser utilizado para comparar empresas de diferentes segmentos da economia, porém, pode ser analisado com base em taxas de outras empresas do mesmo setor (VAN HORNE; WACHOWICZ, 2008).

Os investidores podem calcular o ROE no início ou final de um período, com vistas a analisar a mudança ocorrida ao longo do exercício contábil, como forma de acompanhar o progresso da empresa e promover ajustes em seu planejamento, quando necessário. O índice pode ser expresso por:

$$ROE = \frac{Lucro\ Liquido}{Patrimônio\ Líquido} \tag{2.19}$$

Assaf Neto (2007) pondera que o ROE deve ser comparado com a taxa mínima de retorno exigida pelo acionista, de tal modo que o investimento seja atraente e ofereça uma rentabilidade, pelo menos, igual ao custo de capital.

O custo de capital reflete a taxa de retorno esperada pelo investidor e, devido a sua subjetividade e alto grau de incerteza, é uma variável de difícil mensuração (DONOVAN; NUÑES, 2010).

Por fim, a literatura acadêmica denota a relevância dos índices de rentabilidade abordados nesta seção. Na visão de Assaf Neto (2007) e Wernke (2008), os índices ROA, ROE e ROI revelam diferentes perspectivas sobre o lucro gerado pelas atividades de uma empresa e geram subsídios para análise financeira.

2.5 ANÁLISE DE AGRUPAMENTOS

Segundo Johnson e Wichern (2007), o objetivo da Análise de Agrupamentos ou *clustering* é encontrar o agrupamento ou *cluster* natural de itens ou variáveis, com desenvolvimento de uma determinada escala a fim de medir o grau de associação dos objetos.

A Análise de Agrupamentos consiste em uma técnica estatística que possibilita a criação de agrupamentos de itens diversos, de acordo com as semelhanças apresentadas por esses itens em relação a algum critério de seleção, determinado previamente pelo pesquisador (HAIR *et al.*, 2010).

O objetivo da Análise de Agrupamentos é classificar um pequeno número de grupos que tenham a característica de serem homogêneos internamente, heterogêneos entre si e mutuamente excludentes (HAIR *et al.*, 2010). Portanto, quando plotados geometricamente, objetos dentro de agrupamentos devem estar próximos, enquanto espera-se que os aglomerados estejam tão distantes uns dos outros quanto possível.

A Análise de Agrupamentos é uma técnica distinta dos métodos de reconhecimento de padrões. Agrupar é uma técnica primitiva, no sentido de que nenhuma suposição é feita quanto à estrutura do agrupamento. Por outro lado, nas técnicas de reconhecimento de padrões existem grupos conhecidos, onde o objetivo consiste em alocar uma nova observação em um destes grupos, a partir de variáveis discriminantes.

Hair *et al.* (2010) afirmam que qualquer aplicação da Análise de Agrupamentos deve ter um argumento a respeito de quais variáveis são selecionadas, seja em relação a considerações teóricas e conceituais ou referente a situações práticas. Assim, o pesquisador deve incluir apenas aquelas variáveis que caracterizam os objetos agregados e se relacionam especificamente aos objetivos da Análise de Agrupamentos.

Para a execução do método de Análise de Agrupamentos, os passos a seguir podem ser realizados (PEREIRA, 1999):

1. Cálculo das distâncias entre os objetos estudados no espaço multiplano de todas as variáveis consideradas;
2. Sequência de agrupamento por proximidade geométrica;

3. Reconhecimento dos passos de agrupamento para identificação coerente de grupos dentro do universo de objetos estudados.

Ademais, o planejamento de análise de conglomerados em cinco estágios, proposto por Aldenderfer e Blashfield (1984), adota a seguinte estrutura:

1. Selecionar a amostra;
2. Determinar as variáveis;
3. Definir a medida de similaridade e algoritmo de aglomeração;
4. Delimitar o número de grupos;
5. Validar os resultados.

A escolha das variáveis a serem utilizadas, como a primeira etapa a ser observada na efetivação da técnica de Análise de Agrupamentos, é um dos passos mais importantes do processo de pesquisa, porém, infelizmente, um dos menos compreendidos (ALDENDERFER; BLASHFIELD, 1984).

Os algoritmos mais utilizados se dividem em duas categorias gerais, a saber: hierárquica e não hierárquica. Segundo Hair *et al.* (2010), sempre existirão vantagens e desvantagens na escolha de uma ou outra abordagem, pois os *sotwares* utilizam diferentes algoritmos.

Métodos hierárquicos demandam a especificação do quanto os objetos são diferentes, a fim de identificar diferentes *clusters*. *Softwares* estatísticos podem utilizar medidas de similaridade para estimar a distância entre pares de objetos.

A partir do ponto em que são definidas as variáveis que irão nortear a Análise de Agrupamentos, bem como as medidas de similaridade escolhidas, torna-se necessário decidir sobre o algoritmo que fará o processo de agrupamento, uma vez que a formação dos *clusters* é uma consequência do critério escolhido para medir a distância entre as variáveis definidas e do método de agregação utilizado (FÁVERO *et al.*, 2009).

Um procedimento para executar a validação dos resultados de uma Análise de Agrupamentos consiste no particionamento da amostra original e comparação das soluções obtidas em ambos os casos, verificando a correspondência dos resultados (HAIR *et al*., 2010).

Nas próximas seções serão exploradas as diferentes métricas utilizadas como medidas de distância entre observações multivariadas, bem como os algoritmos usualmente empregados na Análise de Agrupamentos.

2.5.1 Medidas de similaridade

Os esforços envidados para a obtenção de uma estrutura de grupos demandam a elaboração de medidas de proximidade ou similaridade. Frequentemente, há certa parcela de subjetividade envolvendo a escolha da medida de similaridade para o processo de *clustering* e existem aspectos a serem avaliados, como a natureza da variável em estudo, as escalas de medida, e o conhecimento sobre a matéria em questão (JOHNSON; WICHERN, 2007).

Desta forma, a determinação da medida de distância a ser utilizada para avaliar a semelhança entre os objetos e determinar aglomerados pode gerar um efeito relevante sobre a análise e as conclusões subsequentes.

A distância euclidiana é frequentemente a técnica preferida no *clustering* (JOHNSON; WICHERN, 2007). Trata-se da distância geométrica entre observações no espaço multidimensional, expressa por:

$$d(\underline{X},\underline{Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + \left(x_p - y_p\right)^2} \qquad (2.20)$$

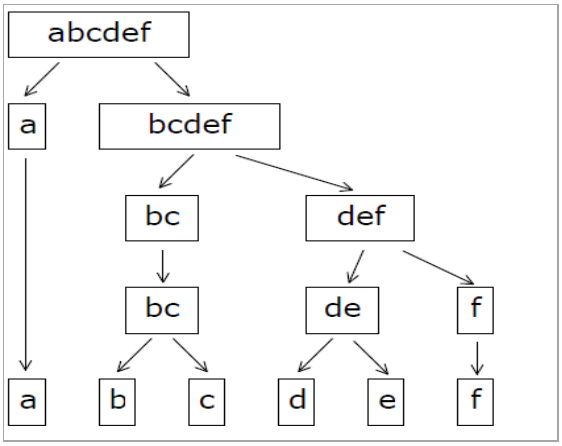
2.5.2 Algoritmos de agrupamento hierárquico

Johnson e Wichern (2007) ponderam a existência de ampla variedade de algoritmos de agrupamentos, cuja finalidade consiste em formar grupos sem necessidade de avaliar todas as possíveis configurações. Neste contexto, técnicas de agrupamento hierárquico processam uma série de junções ou de sucessivas divisões e podem, ainda, ser classificados em algoritmos aglomerativos e divisivos.

No método de agrupamento hierárquico aglomerativo, inicialmente tem-se o número de grupos igual à quantidade de objetos. Primeiramente, as observações mais similares (próximas) são agrupadas e estes grupos iniciais, posteriormente, são agregados de acordo com suas similaridades, até que seja formado um grupo único com todos os objetos.

Nos métodos divisivos, um único grupo de objetos inicial é dividido em dois subgrupos, cujos objetos em um subgrupo encontram-se distantes dos objetos do outro. Estes subgrupos são posteriormente divididos em grupos dissimilares; o processo continua até que exista um número de subgrupos igual ao número de objetos, ou seja, até que cada objeto forme um grupo. A FIGURA 6 apresenta o funcionamento de um algoritmo hierárquico divisivo:

FIGURA 6 - MÉTODO HIERÁRQUICO DIVISIVO



FONTE: O autor (2016).

Jonhson e Wichern (2007) descrevem os seguintes passos a serem observados por algoritmos de agrupamento hierárquico aglomerativo:

1. Inicialmente tem-se *N* grupos, cada um contendo um dos *N* objetos que pretende-se agrupar. Então, calcula-se a matriz simétrica ***D*** de distâncias (ou similaridades) *N x N*, composta de elementos d_{ij} que representam a distância entre o objeto *i* e o objeto *j.* Portanto, tem-se:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \quad (2.21)$$

2. Na matriz *D*, encontra-se o par de grupos que possui menor distância (menor valor de d_{ij}) e agrega-se estes dois grupos *A* e *B* em um novo agrupamento.

3. Une-se os *clusters A* e *B*, nomeando o novo grupo como *(AB) e* atualiza-se as entradas na matriz de distâncias deletando as linhas e colunas correspondentes nos grupos A e B, além de adicionar-se uma linha e uma coluna com base nas distâncias entre o *cluster (AB)* e os grupos remanescentes.
4. Repete-se os passos *2* e *3* por *N-1* vezes, gravando quais grupos deram origem a este novo, bem como as respectivas medidas de distância (ou similaridade), até que seja formado um grupo único.

Os resultados dos métodos hierárquicos podem ser apresentados na forma de um diagrama bidimensional conhecido como dendrograma. Por meio desta ferramenta gráfica e conhecimento prévio sobre os dados, determina-se uma distância de corte para definir os grupos (JOHNSON; WICHERN, 2007).

2.5.3 Métodos de ligação

A seção anterior discorreu sobre o método aglomerativo hierárquico, onde houve referência à maneira pela qual se agrupa objetos similares. Tal agrupamento é realizado por meio de métodos de ligação, os quais serão descritos a seguir, de acordo com Johnson e Wichern (2007).

2.5.3.1 Ligação simples

Nas ligações simples, os grupos são formados a partir dos dois grupos com menor distância. Encontra-se a menor distância da matriz *D*, forma-se o grupo dos objetos correspondentes, por exemplo, o grupo *AB*. No passo 3 do agrupamento hierárquico, a distância entre o grupo *AB* e qualquer outro grupo *C*, é dado por:

$$d_{(AB)C} = min\{ d_{AC}, d_{BC} \} \quad (2.22)$$

Onde:

- d_{AC} = distância entre os grupos *A* e *C;*
- d_{BC} = distância entre os grupos *B* e *C.*

2.5.3.2 Ligação completa

Na ligação completa, o processo é semelhante ao de ligação simples, com a única diferença que a distância entre um determinado grupo *AB* e os outros grupos, é definida de acordo com a máxima distância de cada grupo, ou seja:

$$d_{(AB)C} = max\{ d_{AC}, d_{BC} \} \tag{2.23}$$

Onde:

- d_{AC} = distância entre os grupos *A* e *C;*
- d_{BC} = distância entre os grupos *B* e *C.*

2.5.3.3 Método das médias das distâncias

Neste método, a distância entre dois grupos é dada pela distância média entre os pares de objetos que pertencem a cada grupo. Encontra-se a menor distância entre dois grupos, forma-se o novo grupo a partir destes, e determina-se a distância entre este grupo formado e os outros grupos. Portanto, tem-se:

$$d_{(AB)C} = \frac{\sum_i \sum_j d_{ij}}{N_{(AB)} N_C} \tag{2.24}$$

Onde:

- d_{ij} = distância entre o objeto *i* do grupo *AB* e o objeto *j* no grupo *C;*
- N_{AB} = número de itens do grupo *AB;*
- N_C = número de itens do grupo *C.*

2.5.3.4 Método do centroide

Define a coordenada de cada grupo como sendo a média das coordenadas de seus objetos. Uma vez obtida essa coordenada, denominada centroide, a distância entre os grupos é obtida através do cálculo das distâncias entre os centroides.

2.5.3.5 Método de Ward

Este método possui a premissa de minimizar a perda de informação ao se juntar dois grupos. A perda de informação pode ser vista como o aumento na soma dos quadrados dos desvios (ESS - *Error Sum of Squares*) para um determinado *cluster k*, onde ESS_K representa a soma dos quadrados dos desvios de cada item do aglomerado em relação à média do grupo (centróide).

Se existem *k* grupos, a ESS é definida como a soma de cada ESS_K, ou seja, $ESS = ESS_1 + ESS_2 + ... + ESS_K$.

Em cada passo da análise, a união de possíveis pares de *clusters* é considerada e junta-se os dois grupos que resultem na combinação que corresponda ao menor aumento sobre o ESS. Assim, tem-se:

$$ESS = \sum_{j=1}^{N} \left(\underline{X}_j - \underline{\bar{X}}\right)' \left(\underline{X}_j - \underline{\bar{X}}\right) \tag{2.25}$$

Onde:

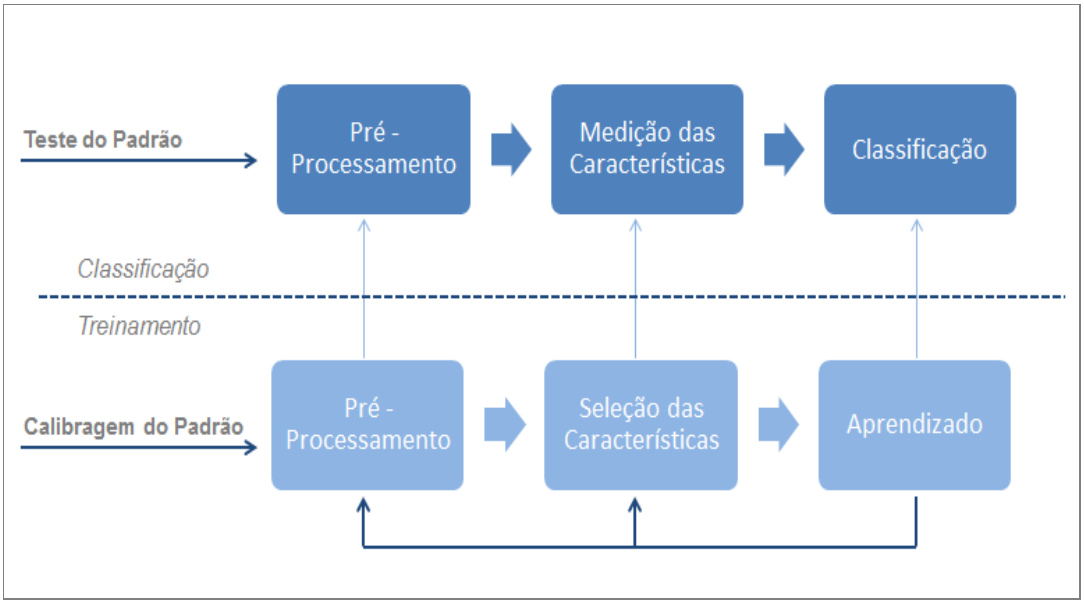
- $\underline{X}_j$ = medida multivariada associada com o *j-ésimo* item;
- $\underline{\bar{X}}$ = média global de todos os itens.

2.6 MÉTODOS ESTATÍSTICOS PARA RECONHECIMENTO DE PADRÕES

Em reconhecimento de padrões com abordagem estatística, um padrão é representado como um conjunto de características, chamado de vetor de características *d-dimensional*.

Os conceitos da teoria de decisão estatística são utilizados para estabelecer fronteiras de decisão entre classes e padrões e o sistema de reconhecimento é operado em dois modos: treinamento e classificação (JAIN; DUIN; MAO, 2000), conforme apresentado na FIGURA 7.

FIGURA 7 - PROCESSO DE RECONHECIMENTO DE PADRÕES



FONTE: Adaptado de JAIN; DUIN; MAO (2000).

O aumento da concorrência em segmentos estratégicos da economia e a crescente pressão para a maximização de resultados podem impulsionar as organizações a investir em modelos de negócios cada vez mais eficientes para expandir seus negócios e, ao mesmo tempo, buscar alternativas para prever resultados financeiros em relação a seus concorrentes.

De tal modo, observa-se a aplicação de modelos estatísticos multivariados para reconhecimento de padrões, baseados em índices econômico-financeiros, tais como modelos para avaliação de risco de crédito. Por meio destas técnicas, as observações são alocadas em grupos pré-existentes, com base em variáveis definidas pelo mecanismo de classificação empregado.

Neste contexto, Scarpel (2005) advoga que, dentre os diferentes métodos quantitativos existentes, as técnicas multivariadas costumam ser largamente empregadas na área de reconhecimento supervisionado de padrões. O autor afirma, ainda, que tais modelos auxiliam o tomador de decisão na classificação de observações futuras, obtendo-se previsões nas quais as observações são enquadradas em uma das classificações existentes.

Segundo Kendall (1980), a análise multivariada aborda as relações entre as variáveis dependentes e os registros analisados, visando à simplificação da estrutura dos dados, classificação, agrupamento de variáveis, análise de interdependência e de dependência, além de formulação e teste de hipóteses.

2.6.1 Análise discriminante

A Análise Discriminante é uma técnica estatística multivariada que busca a separação de conjuntos distintos de objetos e alocação de novas observações em grupos previamente definidos. Segundo Hair *et al.* (2010), o método possui raízes na estatística univariada e sua extensão multivariada introduz conceitos adicionais e questões com particular relevância.

No entendimento de Johnson e Wichern (2007), a Análise Discriminante vislumbra a obtenção de uma combinação linear das características observadas que apresente maior poder de discriminação entre populações. Segundo os autores, a técnica é frequentemente utilizada para definição de regras para designar novos indivíduos aos grupos e para investigação de diferenças observadas, quando os relacionamentos causais não são bem entendidos.

As variáveis que melhor diferenciam os grupos são utilizadas para a criação de funções discriminantes a serem utilizadas para alocar novos objetos ou observações no grupo mais adequado (HAIR *et al.*, 2010).

Para Johnson e Wichern (2007), os objetivos da Análise Discriminante são:

1. Descrever, graficamente ou algebricamente, características diferentes de populações desconhecidas para encontrar discriminantes, cujos valores numéricos sejam tais que, as populações possam ser separadas o tanto quanto possível;
2. Ordenar objetos em dois ou mais grupos previamente definidos. Neste caso, a ênfase se caracteriza em derivar uma regra para otimizar a alocação de novos objetos nas classes definidas.

Marriott (1974) considera a Análise Discriminante como uma extensão da Análise de Variância Multivariada (MANOVA). O autor destaca, porém, que há uma importante diferença em relação aos objetivos a serem alcançados, pois enquanto a MANOVA verifica a existência de diferenças significativas entre grupos, a Análise Discriminante consiste em desenvolver e utilizar funções para classificar objetos em grupos pré-determinados.

Igualmente, a Análise Discriminante também possui semelhanças com a Regressão Múltipla, apesar de finalidades distintas. Neste sentido, a técnica de

Análise Discriminante pressupõe que as variáveis independentes possuem distribuição normal multivariada e utiliza uma estratégia de encontrar processos apurados de classificação de indivíduos (REIS, 2001).

Hair *et al.* (2010) destacam a semelhança entre as técnicas de Análise Discriminante e Regressão Logística, no que tange ao objetivo de identificar o grupo ao qual um determinado objeto pertence. Segundo os autores, a Regressão Logística, na sua forma básica, limita-se à classificação de dois grupos.

2.6.1.1 Pressupostos do modelo

Existem pressupostos a serem observados para utilização da Análise Discriminante, segundo Hair *et al*. (2010) e Fávero *et al*. (2009), os quais sejam:

- *Normalidade multivariada das variáveis explicativas*: caso a normalidade não seja observada, a análise pode ser realizada para fins descritivos. Neste caso, podera ser utilizada a função discriminante linear de Fisher para classificação;
- *Homogeneidade das matrizes de variância e covariância*: assume-se que as matrizes de covariância são homogêneas entre os grupos, o que pode ser testado através da condução de testes estatísticos e de análises gráficas. Desvios menores não impossibilitam a aplicação da técnica, sendo possível uma decisão para prosseguir a análise, apesar da violação do pressuposto;
- *Presença de linearidade das relações*: o método incorpora relações lineares entre as variáveis discriminantes. A linearidade pode ser testada por meio de exame dos diagramas de dispersão;
- *Ausência de problemas relacionados à multicolinearidade das variáveis explicativas*: alta colinearidade entre as variáveis podem tornar a inversão de matrizes demasiadamente instável.

2.6.1.2 Discriminação e classificação

Johnson e Wichern (2007) afirmam que uma boa classificação deve resultar em pequenos erros, isto é, baixa probabilidade de classificação incorreta das

observações. Os autores ponderam que a regra de classificação deve considerar as probabilidades *a priori* e o custo decorrente de má classificação.

Segundo Scarpel (2005), a Análise Discriminante implica na estimativa das densidades de probabilidades específicas das diferentes populações em estudo. Em relação à escolha da função discriminante, a mesma pode depender do conhecimento prévio dos padrões que serão utilizados no processo de classificação ou pode-se optar por utilizar uma forma funcional específica com parâmetros estimados, utilizando o conjunto de treinamento.

A Análise Discriminante deseja separar objetos advindos de populações diferentes e alocar um novo objeto a uma destas classes. Os objetos são separados ou classificados, com base em alguma medida, que associe o vetor de ρ variáveis aleatórias $\underline{X}' = [X_1, X_2, ..., X_p]$.

Para execução de regras de classificação para duas populações, todos os possíveis resultados são divididos em duas regiões R_1 e R_2, de forma que se uma nova observação caia na região R_1, ela é alocada na população Π_1, caso contrário, na população Π_2. Ademais, as regras de classificação não costumam gerar um método isento de erros, visto que há possibilidade de sobreposição entre as regiões R_1 e R_2 (JOHNSON; WICHERN, 2007).

2.6.1.3 Regiões de classificação

Sejam $f_1(\underline{x})$ e $f_2(\underline{x})$ as funções densidade de probabilidade para as populações Π_1 e Π_2 associadas a um vetor aleatório $\underline{X}$ de dimensão p, e um determinado objeto com as medidas $\underline{X}$, que pode ser incluído em Π_1 ou Π_2.

Seja o espaço amostral Ω, R_1 o conjunto de todas as observações de $\underline{X}$ que podem ser classificadas como Π_1, e $R_2 = \Omega - R_1$ os $\underline{X}$ valores restantes de classificados como Π_2. Considerando que todos os objetos podem ser selecionados somente em uma das duas populações, os conjuntos R_1 e R_2 são mutuamente excludentes. A probabilidade condicional $P(2|1)$, de se classificar um objeto como Π_2, quando, de fato, o mesmo é oriundo de Π_1 é igual a:

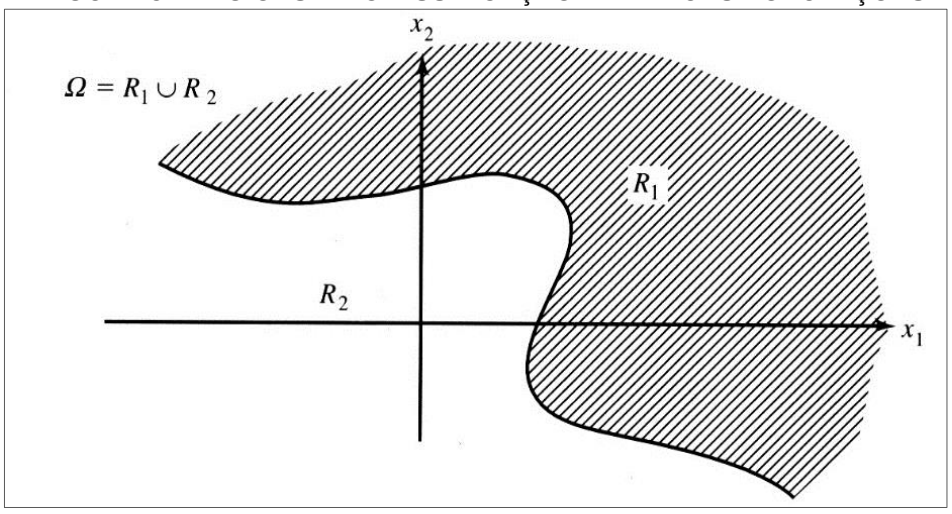
$$P(2|1) = P(\underline{X} \in \Pi_2 | \Pi_1) = \int_{R_2=\Omega-R_1} f_1(\underline{x})\, d\underline{x} \tag{2.26}$$

De forma similar, tem-se a probabilidade condicional *P*(1|2), de se classificar um objeto como Π_1, quando, de fato, o mesmo é oriundo de Π_2:

$$P(1|2) = P(\underline{X} \in \Pi_1 | \Pi_2) = \int_{R_1} f_2(\underline{x})\, d\underline{x} \tag{2.27}$$

A FIGURA 8 ilustra as regiões de classificação R_1 e R_2, para o caso de classificação de *p* = 2 grupos.

FIGURA 8 - REGIÕES DE CLASSIFICAÇÃO PARA DUAS POPULAÇÕES



FONTE: JOHNSON; WICHERN (2007).

A probabilidade de se classificar correta ou incorretamente objetos pode ser descrita como o produto entre as probabilidades *a priori* e as probabilidades condicionais, dadas por:

1. *P(Observação é corretamente classificada como* Π_1)
   = P(Observação é proveniente de Π_1 *e corretamente classificada como* Π_1)

2. *P(Observação é incorretamente classificada como* Π_1)
   = P(Observação é proveniente de Π_2 *e classificada como* Π_1*) = p(1|2)*

3. *P(Observação é corretamente classificada como* Π_2)
   = P(Observação é proveniente de Π_2 *e classificada corretamente como* Π_2*)*

4. *P(Observação é incorretamente classificada como* Π_2)
   = P(Observação é proveniente de Π_1 *e classificada como* Π_2*) = p(2|1)*

De acordo com Johnson e Wichern (2007), a probabilidade p(2|1) representa o volume formado pela f.d.p. $f_1(\underline{x})$ na região R_2.

2.6.1.4 Custo de classificação incorreta (ECM)

Johnson e Wichern (2007) definem a existência de um custo associado à classificação incorreta de objetos. O custo de uma classificação correta é igual à zero, enquanto $c(1|2)$ é o custo de classificação inadequada de uma observação de Π_2 como Π_1 e $c(2|1)$ o custo referente à uma observação de Π_1 incorretamente classificada como Π_2.

Para regras de classificação, o custo médio de classificação incorreta (ECM) pode ser obtido da seguinte forma:

$$ECM = c(2|1)\, p(2|1)\, p_1 + c(1|2)\, p(1|2)\, p_2 \tag{2.28}$$

Onde:

- p_1 = probabilidade *a priori* de se classificar uma observação na população Π_1;
- p_2 = probabilidade *a priori* de se classificar uma observação na população Π_2.

Segundo Johnson e Wichern (2007), as regiões que minimizam o ECM são dadas pelos valores $\underline{x}$ para os quais as seguintes inequações são satisfeitas:

$$R_1 : \left\{ \underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \right\} \tag{2.29}$$

$$R_2 : \left\{ \underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \right\} \tag{2.30}$$

Desta forma, a definição de um ECM mínimo requer (1) a razão de função densidade de probabilidade para uma nova observação $\underline{x}_0$, (2) a razão dos custos de classificação incorreta e (3) a razão das probabilidades *a priori* p_1 e p_2.

2.6.1.5 Classificação de duas populações normais multivariadas com a mesma matriz de covariância

Os procedimentos de classificação baseados em populações normais são bastante comuns em práticas estatísticas, devido a sua simplicidade e alta eficiência (JOHNSON; WICHERN, 2007).

Sejam $f_1(\underline{x})$ e $f_2(\underline{x})$ as funções densidade de probabilidade de populações com distribuições normais multivariadas, sendo o primeiro vetor das médias $\underline{\mu}_1$ e matriz de covariância Σ_1 e o segundo vetor de médias $\underline{\mu}_2$ e matriz de covariância Σ_2. Nesta seção, será abordado o caso especial em que, se as populações possuem a mesma matriz de covariância, resulta-se em um modelo de classificação linear.

Seja a seguinte densidade conjunta do vetor $\underline{X}' = [X_1, X_2, \ldots, X_p]$ para as populações Π_1 e Π_2:

$$f_{i(\underline{x})} = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{1/2}} exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}_i\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}_i\right)\right] \quad i = 1,2 \tag{2.31}$$

Supondo que os parâmetros populacionais $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ sejam conhecidos, após simplificação na razão de $\frac{f_{1(\underline{x})}}{f_{2(\underline{x})}}$, as regiões com mínimo ECM são dadas por:

$$R_1 = \left\{\underline{x} \mid exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}_1\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}_1\right) + \frac{1}{2}\left(\underline{x} - \underline{\mu}_2\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}_2\right)\right] \geq \left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \tag{2.32}$$

$$R_2 = \left\{\underline{x} \mid exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}_1\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}_1\right) + \frac{1}{2}\left(\underline{x} - \underline{\mu}_2\right)' \Sigma^{-1} \left(\underline{x} - \underline{\mu}_2\right)\right] < \left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\} \tag{2.33}$$

Sejam as populações Π_1 e Π_2, descritas pela função densidade de probabilidade normal multivariada acima. Então, a regra de classificação que minimiza o ECM será:

a) Alocar a observação $\underline{X}_0$ em Π_1 se:

$$\left(\underline{\mu}_1 - \underline{\mu}_2\right)' \Sigma^{-1} \underline{X}_0 - \frac{1}{2}\left(\underline{\mu}_1 - \underline{\mu}_2\right)' \Sigma^{-1} \left(\underline{\mu}_1 + \underline{\mu}_2\right) \geq ln\left[\left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (2.34)$$

b) Alocar a observação $\underline{X}_0$ em Π_2, caso contrário.

Na maior parte dos casos práticos, os parâmetros populacionais $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ não são conhecidos, então a regra acima deve ser modificada (JOHNSON; WICHERN, 2007).

Suponha-se que e existam n_1 observações de uma variável aleatória multivariada $\underline{X}' = [X_1, X_2, ..., X_p]$ de Π_1 e n_2 observações de Π_2, com $n_1 + n_2 - 2 \geq p$. Então, a respectiva matriz de dados é dada por:

$$\underline{X}_{1_{(n_1\, x\, p)}} = \begin{bmatrix} x'_{11} \\ x'_{12} \\ \vdots \\ x'_{1n_1} \end{bmatrix} \quad e \quad \underline{X}_{2_{(n_2\, x\, p)}} = \begin{bmatrix} x'_{21} \\ x'_{22} \\ \vdots \\ x'_{2n_2} \end{bmatrix} \quad (2.35)$$

Com base nestas matrizes, os vetores de médias amostrais e matrizes de correlação amostral são dados por:

$$\underline{\bar{x}}_{1_{(p\, x\, 1)}} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \quad e \quad S_{1_{(p\, x\, p)}} = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (x_{1j} - \underline{\bar{x}}_1)(x_{1j} - \underline{\bar{x}}_1)' \quad (2.36)$$

$$\underline{\bar{x}}_{2_{(p\, x\, 1)}} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} \quad e \quad S_{2_{(p\, x\, p)}} = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2} (x_{2j} - \underline{\bar{x}}_2)(x_{2j} - \underline{\bar{x}}_2)' \quad (2.37)$$

Caso seja assumido que as populações possuem a mesma matriz de covariância Σ, as matrizes de covariância amostrais são combinadas e derivam uma única matriz, um estimador sem viés de Σ, calculado por:

$$S_p = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right] S_2 \qquad (2.38)$$

De acordo com Johnson e Wichern (2007), caso não seja observada a normalidade multivariada, deve-se optar por umas das alternativas a seguir:

1. Aplicar alguma transformação sobre os dados e realizar um teste de igualdade de matrizes de covariância de forma a verificar se a regra linear ou a regra quadrática são apropriadas;
2. Utilizar a função linear de Fisher ou a função quadrática, e esperar que o procedimento funcione razoavelmente bem, visto que ambas as regras dispensam o pressuposto de normalidade multivariada dos dados.

Ademais, a base de dados utilizada para classificação pode ser segregada em uma amostra de treinamento e outra de teste. Amostras de treinamento podem ser usadas para desenvolver a função de classificação, enquanto as amostras de teste são empregadas para avaliação de seu desempenho (HAIR *et al*., 2010).

2.6.1.6 Função discriminante linear de Fisher

Fisher desenvolveu uma abordagem para classificação linear, visando transformar observações multivariadas $\underline{X}$ em observações univariadas *y*, de forma que os valores de *y*, derivados das populações Π_1 e Π_2, sejam separados tanto quanto possível.

Para alcançar seu objetivo, Fisher utilizou combinações lineares dos vetores $\underline{X}$, criando-se os termos *y*, visto que combinações lineares podem ser facilmente calculadas. A função linear de Fisher dispensa a premissa de que os dados possuam distribuição normal multivariada, porém, assume a igualdade entre as matrizes de covariância.

Guimarães (2000) afirma que esta função apresenta boas propriedades para a discriminação entre populações com a mesma matriz de covariância, e consiste na ideia básica de criar uma combinação linear das variáveis independentes de forma a definir a variável resposta.

A seguir, será apresentada a regra de classificação de duas populações pela função discriminante linear de Fisher, conforme abordado por Johnson e Wichern (2007). Como os autores destacam que os parâmetros populacionais geralmente não são conhecidos em situações práticas, serão adotados os seus estimadores $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$ e S_p, obtidos de amostras dos grupos Π_1 e Π_2.

De forma geral, a função consiste em selecionar a combinação linear de $\underline{X}$ que alcance a máxima separação entre as médias amostrais univariadas $\bar{Y}_1$ e $\bar{Y}_2$.

Uma combinação linear de $\underline{X}$ assume os valores $y_{11}, y_{12},..., y_{1n1}$ para as observações da primeira população e os valores $y_{21}, y_{22},..., y_{2n2}$ para as observações da segunda população. A separação destes dois conjuntos univariados é analisada pela diferença entre $\bar{Y}_1$ e $\bar{Y}_2$, expressa em unidades de desvio padrão, ou seja:

$$Separação = \frac{[\,\bar{Y}_1 - \bar{Y}_2\,]}{S_y} \tag{2.39}$$

Onde S^2_y é calculador por:

$$S^2{}_y = \frac{\sum_{j=1}^{n_1}\left(y_{1j} - \bar{y}_1\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}_2\right)^2}{n_1 + n_2 - 2} \tag{2.40}$$

Por sua vez, o estimador S_p da matriz $\sum$ é calculado com base nas matrizes de covariância amostrais S_1 e S_2, conforme expressão a seguir:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{2.41}$$

Por fim, a função discriminante linear de Fisher é expressa por:

$$\hat{y} = \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1} \underline{X} \tag{2.42}$$

Esta função maximiza a razão entre o quadrado da distância entre as médias e a variância de y. A seguinte regra de classificação deve ser aplicada para classificação de uma nova observação:

a) Alocar uma nova observação $\underline{X_0}$ na população Π_1 se:

$$\hat{Y}_0 = \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1} \underline{X}_0 \geq \frac{1}{2}\left(\bar{Y}_1 + \bar{Y}_2\right) \tag{2.43}$$

b) Alocar uma nova observação $\underline{X_0}$ na população $\Pi 2$ se:

$$\hat{Y}_0 = \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1} \underline{X}_0 < \frac{1}{2}\left(\bar{Y}_1 + \bar{Y}_2\right) \tag{2.44}$$

2.6.1.7 Critérios para seleção de variáveis discriminantes

Em situações onde a quantidade de variáveis discriminantes pode ser demasiadamente grande, deseja-se selecionar um número relativamente menor de variáveis que contenha tanta informação quanto a coleção original (JOHNSON; WICHERN, 2007).

No entendimento de Hair *et al*. (2010), os resultados de qualquer método de seleção de variáveis devem ser interpretados com cautela, pois não há garantia de que o subconjunto de variáveis selecionado é o melhor, independente do critério de seleção utilizado.

2.6.1.7.1 Lambda de Wilks

Esta métrica expressa a relação entre a variância dentro dos grupos e a variância total observada. Sejam *g* a quantidade de grupos e n_i o tamanho da amostra do *i-ésimo* grupo*, o* Lambda de Wilks (Λ) pode ser expresso pela razão dos determinantes das matrizes $\widehat{H}$ e $\widehat{W}$ (JOHNSON; WICHERN, 2007):

$$\Lambda = \frac{\left|\widehat{W}\right|}{\left|\widehat{H} + \widehat{W}\right|} \tag{2.45}$$

Onde, as matrizes $\widehat{H}$ e $\widehat{W}$ são expressas por:

$$\widehat{H} = \sum_{i=1}^{g} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \quad (2.46)$$

$$\widehat{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{X}_{ij} - \bar{X}_i)\left(\bar{X}_{ij} - \bar{X}_i\right) \quad (2.47)$$

Quanto maior o for o poder discriminatório da variável, menor será o seu índice, sendo os valores situados no intervalo $0 < \Lambda \leq 1$.

Quando a estatística possui um valor igual a 1, significa que todas as médias dos grupos são iguais. Valores próximos de zero indicam baixa variabilidade intragrupos em relação ao total, ou seja, a maior parte da variabilidade é atribuída às diferenças entre as médias dos grupos.

2.6.1.7.2 Traço de Hotelling-Lawley

De acordo com Johnson e Wichern (2007), esta medida é definida como:

$$V = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} W_{ij} * \sum_{k=1}^{g} n_k(\bar{X}_{ik} - \bar{X}_i)\left(\bar{X}_{jk} - \bar{X}_j\right) = tr\left[\widehat{H}\widehat{W}^{-1}\right] \quad (2.48)$$

Onde:

- p = número de variáveis do modelo;
- g = número de grupos;
- n_k = tamanho da amostra no *k-ésimo* grupo;
- $\bar{X}_{ik}$ = média da *i-ésima* variável para o *k-ésimo* grupo;
- $\bar{X}_i$ = média da *i-ésima* variável para todos os grupos combinados;
- $\bar{X}_j$ = média da *j-ésima* variável para todos os grupos combinados;
- W_{ij} = elemento da matriz inversa de covariância intragrupos.

Quanto maior a diferença entre as médias dos grupos, maior o valor da estatística. Portanto, uma maneira de avaliar a contribuição de uma variável é a

averiguação do incremento sobre o Traço de Hotelling-Lawley, quando incluída ao modelo.

2.6.1.7.3 Traço de Pilai

Johnson e Wichern (2007) definem esta medida por:

$$U = tr\left[\frac{\widehat{H}}{\widehat{H}+\widehat{W}}\right] \tag{2.49}$$

Onde, as matrizes $\widehat{H}$ e $\widehat{W}$ são expressas em (2.46) e (2.47).

2.6.1.7.4 Raiz máxima de Roy

Segundo Johnson e Wichern (2007), esta métrica é expressa por:

$$\Theta = \lambda_1 = \sum_{j=1}^{p} \frac{\lambda_j}{1+\lambda_j} \tag{2.50}$$

Onde λ_j são os autovalores obtidos na solução característica, dada por:

$$\left(\widehat{H} - \lambda_j \widehat{W}\right) e_j = 0 \quad com\ j = 1,2,\dots,p \tag{2.51}$$

Onde e_j são os autovetores associados aos autovalores λ_j.

2.6.1.8 Critérios para avaliação da função discriminante

Uma importante maneira de se avaliar o desempenho de qualquer procedimento de classificação é calcular sua taxa de erro ou probabilidade de classificação incorreta (JOHNSON; WICHERN, 2007).

Dentre as ferramentas de analisar uma classificação, a Taxa de Erro de Reconhecimento (*TPM - Total Probability of Misclassifications*) é dada por:

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \qquad (2.52)$$

Onde p_1 e p_2 são as probabilidades de uma observação pertencer às populações Π_1 e Π_2, respectivamente.

O valor que minimiza a TPM, denominado de Taxa Ótima de Erro (OER - *Optimum Error Rate*), é obtido pela escolha adequada das regiões R_1 e R_2, determinadas por:

$$R_1 : \left\{ \underline{x} \,/\, \frac{f_2(\underline{x})}{f_1(\underline{x})} \geq \frac{p_2}{p_1} \right\} \quad e \quad R_2 : \left\{ \underline{x} \,/\, \frac{f_2(\underline{x})}{f_1(\underline{x})} < \frac{p_2}{p_1} \right\} \qquad (2.53)$$

Em geral não se conhece a função de distribuição das populações (JOHNSON; WICHERN 2007), portanto, a taxa de erro passa a ser associada à função de classificação amostral, cuja performance pode ser avaliada pelo cálculo da Taxa Real de Erro (*AER - Actual Error Rate*):

$$AER = p_1 \int_{\hat{R}_2} f_1(\underline{x})dx + p_2 \int_{\hat{R}_1} f_2(\underline{x})dx \qquad (2.54)$$

Onde $\hat{R}_1$ e $\hat{R}_2$ representam a região de classificação determinada pelo tamanho das amostras n_1 e n_2, respectivamente.

A abordagem de Lachenbruch também pode ser utilizada para avaliação da eficiência da regra de classificação, com base nas seguintes etapas:

a) Escolher um dos grupos (amostras);
b) Descartar uma observação do grupo;
c) Construir uma função discriminante para as (n_1 – *1)* observações restantes do grupo escolhido e para as n_2 observações do segundo grupo, ou seja, para ($n_1 + n_2 - 1$*)* observações;
d) Classificar a observação descartada com a função obtida anteriormente;

e) Realocar a observação descartada e repetir os passos "*a"* e "*b"* para todas as observações do primeiro grupo;
f) Repetir os passos anteriores para o segundo grupo;
g) Finalmente, ajustar a função discriminante para o total das $n = n_1 + n_2$ observações.

Desta forma, obtêm-se as seguintes funções de probabilidade:

$$P(2|1) = \int_{R_2} f_1(\underline{x})dx = \frac{n_{1/2}}{n_1} \quad e \quad P(1|2) = \int_{R_1} f_2(\underline{x})dx = \frac{n_{2/1}}{n_2} \tag{2.55}$$

Onde:

- P(2|1) = probabilidade de classificação de um objeto em Π_2, quando pertence a Π_1;
- *P*(1|2) = probabilidade de classificação de um objeto em Π_1, quando pertence a Π_2;
- $n_{2/1}$ = número de itens de Π_2 classificados incorretamente como de Π_1;
- $n_{1/2}$ = número de itens de Π_1 classificados incorretamente como de Π_2.

Por sua vez, a proporção total esperada de acertos será dada por:

$$\hat{E}(APER) = \frac{n_{1/2} + n_{2/1}}{n_1 + n_2} \tag{2.56}$$

2.6.2 Regressão logística

Os métodos de regressão têm como objetivo principal descrever a relação entre uma variável resposta e uma ou mais variáveis explicativas. Segundo a especificação do modelo clássico de Regressão Linear Múltipla, o comportamento de uma variável dependente é uma função de um conjunto de variáveis independentes.

A Regressão Logística diferencia-se das técnicas de Regressão Linear devido ao fato de a variável resposta ser binária ou dicotômica. Apesar de haver

diferenças no que se refere a hipóteses e estimação de parâmetros, o modelo logístico se baseia em princípios que norteiam os modelos de Regressão Linear (HOSMER; LEMESHOW, 2000).

Na Regressão Logística, não existem restrições com relação à normalidade multivariada na distribuição das variáveis independentes e à igualdade das matrizes de covariâncias dos dois grupos, tal como ocorre na Análise Discriminante. Ademais, a Regressão Logística representa uma abordagem de classificação nos casos em que as variáveis em estudo são qualitativas (JOHNSON; WICHERN, 2007).

Para que seja estabelecida uma relação linear entre as variáveis independentes é necessária a aplicação de uma transformação sobre a variável resposta, como *logit*, *probit* e *complemento log-log*.

2.6.2.1 O modelo logístico

Segundo Hosmer e Lemeshow (2000), dado o modelo de Regressão Linear Simples:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad para\ i = 1,2,\dots,n \tag{2.57}$$

Onde:

- Y_i = variável resposta*;*
- β_0 = intercepto*;*
- β_1 = coeficiente angular (*slope*)*;*
- X_i = variável explicativa;
- ϵ_i = resíduos do modelo, os quais possuem distribuição de probabilidade Normal (μ;σ^2), variância constante (homocedasticidade) e independência em relação à variável explicativa *Xi*.

O valor esperado para a variável resposta *Yi* é dado por:

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad para\ i = 1,2,\dots,n \tag{2.58}$$

Considerando que a variável resposta Y_i seja dicotômica, esta será uma variável aleatória com distribuição *Bernoulli*, expressa por:

$$\begin{cases} P(Y_i = 1) = \pi_i & se \quad Y_i = 1 \\ P(Y_i = 0) = 1 - \pi_i & se \quad Y_i = 0 \end{cases} \tag{2.59}$$

Pela definição de valor esperado, tem-se que:

$$E(Y_i) = \pi_i \tag{2.60}$$

Substituindo a equação (2.60) em (2.58), obtém-se a seguinte equação:

$$\pi_i = \beta_0 + \beta_1 X_i \tag{2.61}$$

Na visão de Hosmer e Lemeshow (2000), existem alguns problemas na utilização do modelo de Regressão Linear quando a variável resposta é binária:

1. Os erros não possuem distribuição normal: cada resíduo ϵ_i do modelo somente poderá assumir os seguintes valores:

$$\begin{cases} \epsilon_i = 1 - \beta_0 - \beta_1 X_i & se \quad Y_i = 1 \\ \epsilon_i = -\beta_0 - \beta_1 X_i & se \quad Y_i = 0 \end{cases} \tag{2.62}$$

2. Variâncias heterogêneas: a variância de Y_i para o modelo de Regressão Linear Simples é expressa por:

$$\sigma^2(Y_i) = \pi_i(1 - \pi_i) = E(Y_i)\big(1 - E(Y_i)\big) \tag{2.63}$$

Tendo em vista que:

$$\epsilon_i = Y_i - \pi_i \tag{2.64}$$

Como o valor de π_i é constante, observa-se que:

$$\sigma^2(\epsilon_i) = \pi_i(1-\pi_i) = (\beta_0 + \beta_1 X_i) + (1 - \beta_0 - \beta_1 X_i) \quad (2.65)$$

Assim, pode se verificar que o desvio padrão dos resíduos depende de X_i e, portanto, há heterocedasticidade (variância não constante).

3. Restrição na função resposta: como a função representa as probabilidades para a variável resposta binária, tem-se que:

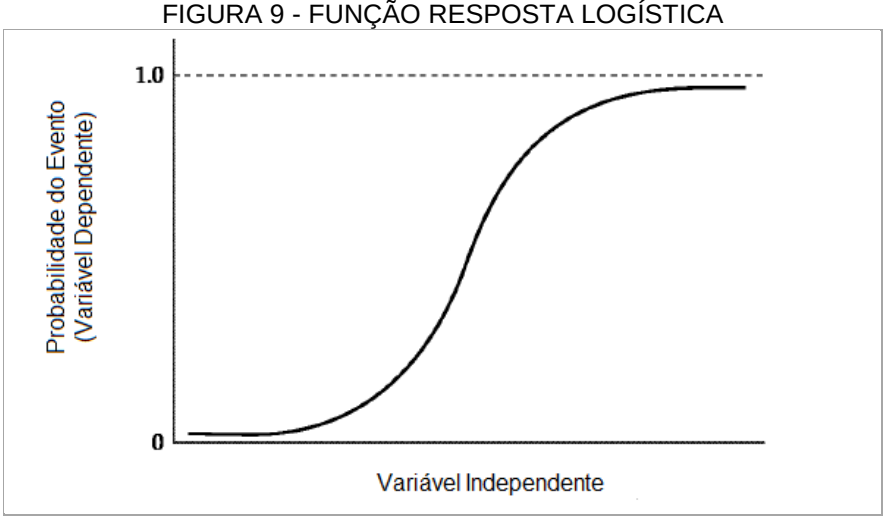
$$0 \leq E(Y_i) = \pi_i \leq 1 \quad (2.66)$$

Johnson e Wichern (2007) reiteram inconvenientes em se utilizar um modelo de Regressão Linear no caso de variável resposta dicotômica:

1. Os valores estimados para a variável resposta *Y* pode ser maior que 1 ou menor que 0, devido à formulação do valor esperado no modelo linear;
2. Uma das hipóteses para a análise de regressão, a variância da variável resposta constante para todos os valores das variáveis independentes, não seria observada neste caso.

Segundo os autores, uma alternativa para tal problema consiste na utilização de um modelo de Regressão Logística, onde as probabilidades 0 e 1 são encontradas assintoticamente, conforme apresentado na FIGURA 9:

FIGURA 9 - FUNÇÃO RESPOSTA LOGÍSTICA



FONTE: Adaptado de HAIR *et al.*(2010).

A Regressão Logística é um modelo de regressão não linear com a seguinte formulação (HOSMER; LEMESHOW, 2000):

$$P_i = E\left(\frac{y_i}{x_i}\right) = \frac{e^{\beta_1+\beta_2\, x_{2i}\, +\cdots+\beta_k\, x_{ki}}}{1 + e^{\beta_1+\beta_2\, x_{2i}\, +\cdots+\beta_k\, x_{ki}}} \tag{2.67}$$

De igual modo, a função P_i pode ser expressa por:

$$P_i = E\left(\frac{y_i}{x_i}\right) = \frac{e^{z_i}}{1 + e^{z_i}} \tag{2.68}$$

Onde:

$$z_i = \beta_1 + \beta_2\, x_{2i}\, + \cdots + \beta_k\, x_{ki} \tag{2.69}$$

Desta forma, o modelo logístico é crescente sem assumir valores fora do intervalo (0,1), conforme ilustrado na FIGURA 9. Vale ressaltar, ainda, a possibilidade de o modelo poder ser linearizado e, portanto:

$$1 - P_i = \frac{1}{1+e^{z_i}} \tag{2.70}$$

Assim, observa-se que:

$$\frac{P_i}{1 - P_i} = e^{z_i} \tag{2.71}$$

O quociente $\left(\frac{P_i}{1-P_i}\right)$ é conhecido como o "*odds*" (chance). Ao se aplicar o logaritmo neperiano à expressão (2.71) e adicionar a componente residual, obtém-se um modelo de Regressão Logística linearizado:

$$L_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \tag{2.72}$$

Desta forma, tem-se que:

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) \tag{2.73}$$

Onde, o vetor $\underline{\beta}_{(px1)}$ com os parâmetros do modelo e o vetor $\underline{X}$ de variáveis preditoras, são expressos por:

$$\underline{\beta}_{(px1)} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \quad \underline{X}_{(px1)} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ . \\ . \\ X_{(p-1)} \end{bmatrix} \quad \underline{X_i}_{(px1)} = \begin{bmatrix} 1 \\ X_{i,1} \\ X_{i,2} \\ . \\ . \\ X_{i,(p-1)} \end{bmatrix} \tag{2.74}$$

A transformação evidenciada em (2.72) resolve as principais dificuldades do modelo de probabilidade linear. De tal modo, esta transformação é muito importante devido ao fato de se obter muitas propriedades desejáveis do modelo de Regressão Linear, pois o modelo logístico é linear em seus parâmetros e P_i ϵ (0,1), que decorre da definição de probabilidade (HOSMER; LEMESHOW, 2000).

2.6.2.2 Estimação dos parâmetros

Johnson e Wichern (2007) não recomendam a estimação dos parâmetros da Regressão Logística pelo método dos Mínimos Quadrados Ordinários, uma vez que o mesmo incorre em erros heterocedásticos e na possibilidade de a função assumir valores sem significado, tornando impossível a estimação do modelo.

Por esta razão, o modelo de Regressão Logística costuma ser estimado pelo método de máxima verossimilhança (HOSMER; LEMESHOW, 2000). Caso o problema fosse apenas a heterocedasticidade do modelo, a situação poderia ser solucionada pela sua transformação em um modelo de regressão clássica pelo método dos Mínimos Quadrados Ponderados.

A fórmula abaixo mostra a função de máxima verossimilhança a ser aplicada no modelo (HOSMER; LEMESHOW, 2000):

$$L = \prod_{i=1}^{n} f(y_i) \tag{2.75}$$

Onde*:*

- *n* = número de observações;
- $f(y_i)$ = função densidade de probabilidade de y_i.

Desta forma, pode-se expressar a função de verossimilhança por:

$$L = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{2.76}$$

Substituindo P_i pela função de distribuição logística, tem-se:

$$L(\underline{\beta}) = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-x_i\underline{\beta}}}\right)^{y_i} \left(\frac{e^{-x_i\underline{\beta}}}{1 + e^{-x_i\underline{\beta}}}\right)^{1-y_i} \tag{2.77}$$

De modo mais simplificado, pode se escrever:

$$L\left(\underline{\beta}\right) = \prod_{i=1}^{n} \lambda\left(\underline{X_i}\underline{\beta}\right)^{y_i} [1 - \lambda\left(\underline{X_i}\underline{\beta}\right)^{1-y_i}] \tag{2.78}$$

Onde:

- $\underline{X_i}$ = vetor de observações das *k* variáveis explicativas da *i-ésima* observação*;*
- $\underline{\beta}$ = vetor dos *k* parâmetros a estimar*;*
- $\lambda\left(\underline{X_i}\underline{\beta}\right)$ = função de distribuição logística.

A maximização desta função é um problema equivalente à maximização do seu logaritmo, já que a função logaritmo é monótona crescente. Segundo Hosmer e Lemeshow (2000), para facilitar a maximização da função, tem-se o logaritmo da função de verossimilhança, ou função log-verossimilhança, dado por:

$$ln\left(L\left(\underline{\beta}\right)\right) = l\left(\underline{\beta}\right) = \sum_{i=1}^{n} y_i \ln\left(\lambda\left(\underline{X_i}\underline{\beta}\right)\right) + \sum_{i=1}^{n} (1 - y_i) \ln\left(1 - \lambda\left(\underline{X_i}\underline{\beta}\right)\right) \tag{2.79}$$

O estimador de máxima verossimilhança dos *k* componentes de $\underline{\beta}$ corresponde aos valores desses parâmetros que maximizam a função $L(\underline{\beta})$. Para obter este máximo, torna-se necessário calcular a primeira e a segunda derivadas da função. Entretanto, não é possível encontrar diretamente uma solução para este problema que assegure a condição necessária para o máximo da função de verossimilhança (HOSMER; LEMESHOW, 2000).

De acordo com Johnson e Wichern (2007), os valores dos parâmetros que maximizam a função de máxima verossimilhança não possuem uma solução exata, tal como nos modelos de regressão linear clássicos. Os autores ponderam que os parâmetros devem ser determinados numericamente, iniciando-se com valores arbitrários e, posteriormente, executando-se iterações até que a função de verossimilhança seja maximizada.

2.6.2.3 Intervalo de confiança para os parâmetros

Segundo Johnson e Wichern (2007), quando o tamanho da amostra for relativamente grande, uma estimativa $\underline{\hat{\beta}}$ para o vetor de parâmetros é aproximadamente normal com média $\underline{\beta}$ e matriz de covariância, dada por:

$$\widehat{Cov}\left(\underline{\hat{\beta}}\right) \approx \left[\sum_{j=1}^{n} \hat{p}(z_i)\left(1-\hat{p}(z_i)\right) z_i z_i'\right]^{-1} \tag{2.80}$$

A raiz quadrada dos elementos da diagonal desta matriz são os maiores erros padrões dos estimadores $\hat{\beta}_0$, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$, respectivamente. Portanto, o intervalo de confiança de 95% para $\underline{\hat{\beta}_k}$ é dado por:

$$\widehat{\underline{\beta_k}} \pm 1{,}96 * Desvio\ Padrão\left(\underline{\hat{\beta}_k}\right),\ k = 0{,}1, \ldots, n \tag{2.81}$$

O intervalo de confiança pode ser utilizado para julgar a significância dos termos individuais no modelo de Regressão Logística.

2.6.2.4 Classificação de duas populações

Uma vez desenvolvida a função de Regressão Logística, e realizados os devidos testes para cada uma das duas populações, pode-se proceder à classificação de novas observações em grupos (JOHNSON; WICHERN, 2007).

Seja a variável resposta *Y* igual a 1, caso a unidade observacional pertença à população Π_1, e igual 0 se a observação pertence à população Π_2. A seguinte regra de classificação deve ser aplicada para classificar uma nova observação $\underline{Z}$:

a) Classificar a nova observação $\underline{Z}$ na população Π_1, caso o *odds ratio* seja maior que 1, ou seja:

$$\left[\frac{\hat{p}(\underline{Z})}{1-\hat{p}(\underline{Z})}\right] = exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_r X_r) > 1 \qquad (2.82)$$

b) Classificar a nova observação $\underline{Z}$ na população Π_2, caso contrário.

Ademais, Johnson e Wichern (2007) estabelecem outra regra de classificação, relacionando a regra de classificação do modelo logístico com a classificação baseada na função discriminante linear:

a) Classificar a nova observação padronizada $\underline{Z}$ na população Π_1 se a função discriminante linear é maior que 0, ou seja:

$$ln\left[\frac{\hat{p}(\underline{Z})}{1-\hat{p}(\underline{Z})}\right] = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r > 0 \qquad (2.83)$$

b) Classificar a nova observação $\underline{Z}$ na população Π_2, caso contrário.

2.6.2.5 Testes de significância

Após a obtenção dos coeficientes do modelo, deve se avaliar a qualidade da estimação destes parâmetros, a fim de se verificar se as variáveis explicativas são significativas para explicar o comportamento da variável resposta.

Os testes mais utilizados para aferir a qualidade do ajustamento do modelo e a significância individual de seus parâmetros são o teste de Razão de Verossimilhança, o teste de Wald e o teste de *Scores*. A finalidade destes testes de significância, em termos práticos, consiste em testar a hipótese nula de que os coeficientes de regressão β_k são iguais à zero. Os coeficientes que não forem aceitos nos teste podem ser retirados do modelo (HOSMER; LEMESHOW, 2000).

2.6.2.5.1 Razão de verossimilhança

O teste da razão de verossimilhança guarda relação com o conceito do teste *F de Snedecor* aplicado no modelo clássico de Regressão Linear. Na Regressão Logística, o teste é baseado nas diferenças entre os logaritmos da função verossimilhança para os modelos com e sem restrições.

Segundo Hosmer e Lemeshow (2000), os estimadores de máxima verossimilhança maximizam a função log-verossimilhança, portanto, ao se retirar as variáveis o resultado é um valor pequeno para a log-verossimilhança, como ocorre com o R^2 no modelo de regressão clássica. A estatística-teste a ser utilizada é expressa por (HOSMER; LEMESHOW, 2000):

$$\lambda = l\left(\underline{\hat{\beta}}_R\right) - l\left(\underline{\hat{\beta}}_U\right) \tag{2.84}$$

Onde:

- $l\left(\underline{\hat{\beta}}_R\right)$ = valor máximo do logaritmo de log-verossimilhança com os parâmetros iguais a zero;
- $l\left(\underline{\hat{\beta}}_U\right)$ = valor máximo do logaritmo de log-verossimilhança sem restrições.

A hipótese nula, de que os parâmetros são iguais a zero, será rejeitada quando o *p-valor* atingir um determinado limiar estabelecido no teste de hipóteses. Caso contrário, a informação das variáveis independentes permite previsões estatisticamente válidas.

2.6.2.5.2 Teste de Wald

O teste de Wald compara a estimativa de máxima verossimilhança de determinado coeficiente com a estimativa do seu erro padrão. A estatística-teste é calculada por (HOSMER; LEMESHOW, 2000):

$$w_j = \frac{\hat{\beta}_j}{\sqrt{VAR(\hat{\beta}_j)}} \tag{2.85}$$

Onde:

- $\sqrt{VAR(\hat{\beta}_j)}$ = desvio padrão estimado do estimador do parâmetro β_j;
- w_j = estatística com distribuição *Qui-Quadrado e n* graus de liberdade.

Os valores críticos α_j para as estimativas dos parâmetros são os valores para os quais, se o valor do teste de Wald calculado para um determinado β_j for maior que α_j, se rejeita a hipótese nula para um dado nível significância.

2.6.2.6 Medidas de associação múltipla

Os coeficientes de determinação calculados no modelo de regressão clássica não são aplicáveis no modelo logístico, pois a variável dependente pode assumir apenas dois valores (variável dicotômica). Desta forma, foram desenvolvidas algumas técnicas para se avaliar o nível de aderência entre as variáveis explicativas e a variável resposta do modelo (HAIR *et al.,* 2010).

2.6.2.6.1 Pseudo R^2 de McFadden

A proposta desta estatística é uma transformação na razão de verossimilhança, em um paralelo com o R^2 da Regressão Linear clássica. Seu cálculo é expresso por:

$$\rho^2 = R^2{}_{McF} = 1 - \frac{l(\hat{\underline{\beta}}_R)}{l(\hat{\underline{\beta}}_U)} \quad (2.86)$$

Onde:

- $\left(\hat{\underline{\beta}}_R\right)$ = *v*alor máximo do logaritmo de log-verossimilhança com os parâmetros iguais a zero;
- $l\left(\hat{\underline{\beta}}_U\right)$ = valor máximo do logaritmo de log-verossimilhança sem restrições.

Valores mais elevados deste coeficiente estão relacionados à maior capacidade explicativa do modelo. Valores entre 0,2 e 0,4 são considerados satisfatórios (HOSMER; LEMESHOW, 2000).

2.6.2.6.2 R² De Cox e Snell

Segundo Hosmer e Lemeshow (2000), esta métrica é calculada com base no logaritmo da função máxima verossimilhança, e leva em consideração a dimensão da amostra, sendo expressa por:

$$R^2{}_{CS} = 1 - e^{\frac{2}{n} l\left(\hat{\underline{\beta}}_R\right) - l\left(\hat{\underline{\beta}}_U\right)} \quad (2.87)$$

Onde:

- $\left(\hat{\underline{\beta}}_R\right)$ = *v*alor máximo do logaritmo de log-verossimilhança com os parâmetros iguais a zero;
- $l\left(\hat{\underline{\beta}}_U\right)$ = valor máximo do logaritmo da função de máxima verossimilhança sem restrições.

2.6.2.6.3 R² De Nagelkerke

Esta estatística é derivada do coeficiente $R^2{}_{CS}$ podendo, porém, atingir o valor máximo de 1,00. Este indicador é calculado por:

$$R^2{}_N = \frac{R^2{}_{CS}}{R^2{}_{max}} \tag{2.88}$$

Onde*:*

- $R^2{}_N$ = R² de Nagelkerke;
- $R^2{}_{CS}$ = R² de Cox e Snell;
- $R^2{}_{max}$ = valor máximo de $R^2{}_{CS}$, ou seja, seu valor quando $l(\hat{\beta}_U)$ = 0*.*

Valores de $R^2{}_N$ acima de 0,3 são considerados como indicadores de boa qualidade de ajustamento do modelo (HOSMER; LEMESHOW, 2000).

2.6.2.7 Análise de resíduos

Na visão de Hair *et al.* (2010), a análise de resíduos desempenha um importante papel na Regressão Logística, pois identifica as observações para os quais o modelo tem pouca aderência e eventuais *outliers*. Neste sentido, os autores destacam algumas medidas para este tipo de análise, tais como a distância de Cook e a distância de Mahalanobis.

Neste contexto, o resíduo de Pearson é a diferença para cada observação entre o valor observado e a probabilidade estimada dividida pelo desvio-padrão binomial da probabilidade estimada (HAIR *et al.*, 2010), ou seja:

$$e_i = \frac{\delta_i - p_i}{\sqrt{p_i * (1 - p_i)}} \tag{2.89}$$

Onde:

- δ_i= valor observado na variável resposta;
- p_i = probabilidade estimada pelo modelo de Regressão Logística;
- $\sqrt{p_i * (1 - p_i)}$ = desvio padrão de uma distribuição binomial da probabilidade estimada pelo modelo.

Para grandes amostras, o resíduo de Pearson segue uma distribuição normal com desvio padrão igual a 1. Valores absolutos elevados indicam que o modelo não tem aderência à observação em particular.

2.6.3 Medidas de avaliação e comparação de métodos para reconhecimento de padrões

A comparação entre o poder preditivo de métodos multivariados para reconhecimento de padrões pode ser realizada a partir de medidas de avaliação disponíveis na literatura acadêmica.

Um modo de avaliar a eficiência de modelos preditivos consiste em verificar a taxa de acerto obtida na alocação das observações nos grupos pré-definidos, com base na regra de classificação desenvolvida. Hair *et al.* (2010) sugerem a utilização do método *cross validation*, que consiste em segregar os dados em dois grupos:

1. Amostra de treinamento: desenvolvimento da função de classificação;
2. Amostra de teste: utilizada para testar a acurácia da função discriminante ao classificar novas observações.

A Matriz de Confusão, apresentada na TABELA 3, expõe a real situação das observações nos grupos, comparando-a com o reconhecimento apresentado pelo modelo encontrado (JOHNSON; WICHERN, 2007).

Esta ferramenta possibilita o cálculo do percentual de acertos do modelo ao classificar os elementos da amostra de teste, com base na função de classificação derivada da amostra de treinamento.

TABELA 3 - MATRIZ DE CONFUSÃO

		Classificação Prevista	
		Π_1	Π_2
Classificação Real	Π_1	$n_{1/1}$	$n_{1/2}$
	Π_2	$n_{2/1}$	$n_{2/2}$

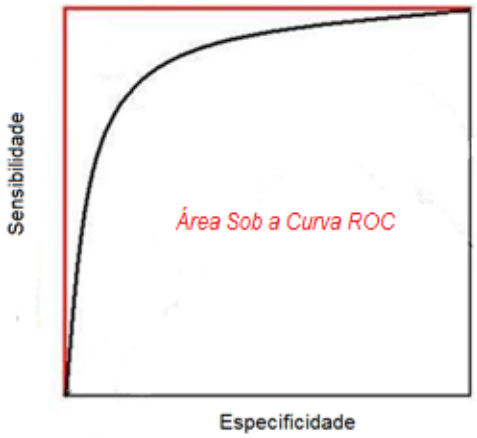
FONTE: Adaptado de JOHNSON; WICHERN (2007).

Onde:

- $n_{1/1}$ = número de itens de Π_1 classificados corretamente como de Π_1;
- $n_{2/1}$ = número de itens de Π_2 classificados incorretamente como de Π_1;
- $n_{2/2}$ = número de itens de Π_2 classificados corretamente como de Π_2;
- $n_{1/2}$ = número de itens de Π_1 classificados incorretamente como de Π_2.

De igual modo, a curva ROC *(Receiver Operating Characteristic*), apresentada na FIGURA 10, permite avaliar o desempenho de um modelo, por meio de um gráfico que indica a variação da sensibilidade e da especificidade, para diferentes pontos de corte (HOSMER; LEMESHOW, 2000).

FIGURA 10 - CURVA ROC



FONTE: Adaptado de HOSMER; LEMESHOW (2000).

A área abaixo da curva de ROC gera uma medida da capacidade do modelo em discriminar as observações em dois grupos. Deste modo, no eixo das ordenadas está representada a sensibilidade do modelo, isto é, a capacidade do modelo em prever os “verdadeiros positivos”. No eixo das abscissas encontra-se a especificidade do modelo, ou seja, sua capacidade em não cometer erros ao identificar os “verdadeiros negativos”.

Segundo Hosmer e Lemeshow (2000), quanto maior for a sensibilidade para valores elevados da especificidade, melhor será o modelo estimado. Neste sentido, uma medida numérica da precisão pode ser obtida pela área da curva, onde o valor igual a 1 sugere um modelo perfeito, enquanto valores próximos de 0,5 indicam baixa capacidade de aderência do modelo.

2.7 TRABALHOS CORRELATOS

De forma a analisar os trabalhos correlatos à esta pesquisa, optou-se pela estruturação de uma bibliometria. Foram revisadas publicações internacionais atinentes ao tema na área de Engenharia, bem como avaliados os principais assuntos, métodos e conteúdo de cada trabalho.

A primeira etapa do procedimento se refere à extração do acervo disponibilizado pela plataforma *Web of Science* ®, onde podem ser localizados importantes periódicos oriundos de áreas de conhecimento diversas. O levantamento foi feito com base na seguinte estruturação de palavras-chave:

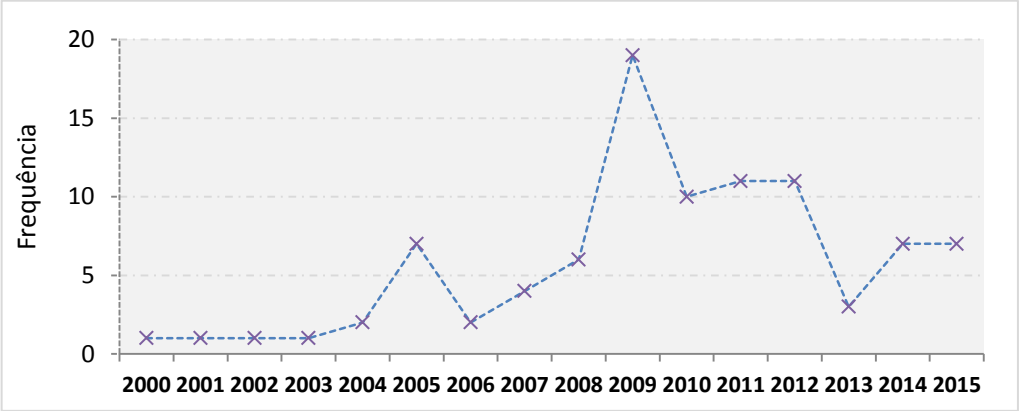
- *('ROE' OR 'Financial Ratios') AND ('Pattern Recognition' OR 'Multivariate') AND ('Insolvency' OR 'Profitability' OR 'Bankruptcy').*

Foram selecionados somente artigos publicados na área de Engenharia, totalizando 95 publicações internacionais nos anos de 2000 a 2015. Por fim, foram realizadas análises quantitativas dos trabalhos para proporcionar discussões e uma visão compartimentada sobre a evolução dos métodos propostos ao longo dos anos.

2.7.1 Aspectos gerais observados na literatura acadêmica

Com a finalidade de explorar o contexto no qual se insere esta pesquisa, foi analisado o resumo (*abstract*) de cada trabalho selecionado. O GRÁFICO 2 apresenta a quantidade de publicações, por ano, correlatas à presente pesquisa.

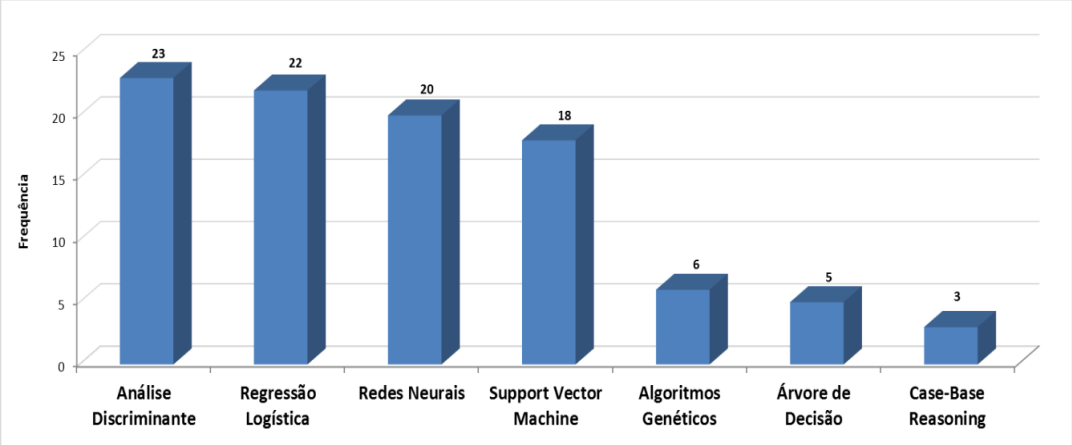
GRÁFICO 2 - FREQUÊNCIA DE TRABALHOS CORRELATOS



FONTE: O autor (2016).

O volume de trabalhos publicados a partir dos anos 2000 denota a relevância acadêmica dos problemas de análise financeira na área de Engenharia com aplicação de modelos preditivos. Ademais, a evolução tecnológica ao longo dos últimos anos possibilita a aplicação de modelos sofisticados para atender problemas demandados pela indústria. Observa-se no GRÁFICO 3, os principais modelos abordados nos artigos selecionados.

GRÁFICO 3 - TÉCNICAS ABORDADAS PELOS TRABALHOS CORRELATOS

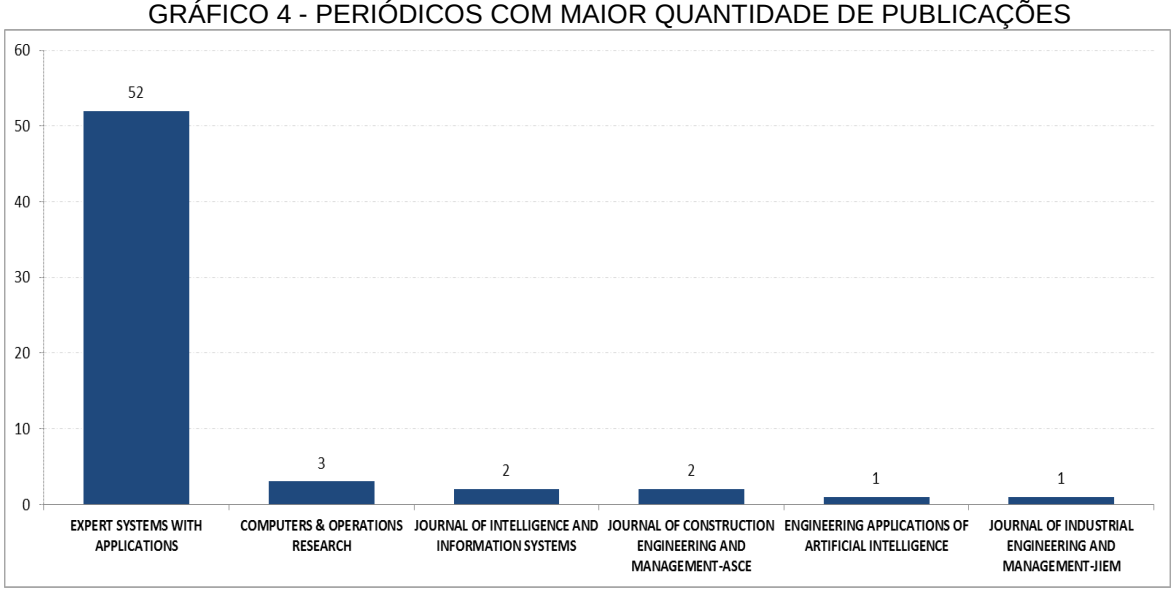


FONTE: O autor (2016).

Notou-se a utilização dos métodos multivariados de Análise Discriminante e Regressão Logística para resolução dos problemas envolvendo previsão de insolvência.

No GRÁFICO 4 são apresentados os periódicos com maior número de publicações até o ano de 2015. Ressalta-se que todos os trabalhos selecionados foram publicados na área de Engenharia.

GRÁFICO 4 - PERIÓDICOS COM MAIOR QUANTIDADE DE PUBLICAÇÕES



FONTE: O autor (2016).

Os periódicos *Expert Systems With Applications, Computers and Operations Research e Engineering Applications of Artificial Intelligence* possuem conceito "*A2*"

no Qualis da área de Engenharias III, de acordo com a classificação disponibilizada pela Comissão Avaliadora da CAPES.

2.7.2 Descrição sumária de trabalhos correlatos

Nesta seção, serão brevemente abordadas as contribuições acadêmicas de alguns trabalhos alinhados aos objetivos desta pesquisa.

Ahn *et al*. (2000) propõem um sistema de inteligência híbrido. O modelo foi construído com a finalidade de prever a falência de empresas com base em informações financeiras passadas, combinando a *Rough Set Approach* com o método de Redes Neurais. Segundo os autores, foi possível a utilização de um menor número de índices financeiros e variáveis qualitativas, na medida em que não houve perda de informação.

Em seu trabalho, Atyia (2001) apresentou contribuições para o problema de previsão de falência de empresas e análise de rentabilidade. Primeiramente, foi desenvolvido um modelo de Redes Neurais com a proposta de indicadores inovadores, inspirado pelo tradicional modelo de risco de crédito de Merton. Foi mostrado que a utilização destes indicadores, em adição aos tradicionais índices financeiros proporcionou um significativo aumento da acurácia da previsão de 81,46% para 85,5% para um horizonte de tempo de três anos.

O artigo de Shin *et al*. (2005), visa investigar a eficácia da aplicação do método *Support Vector Machine* (SVM) para o problema de previsão de falência. Segundo os autores, apesar de os modelos de Redes Neurais apresentarem bom desempenho em tarefas de reconhecimento de padrões, demandam uma tarefa árdua de se encontrar a estrutura adequada e uma solução ótima, bem como a calibragem da rede. Assim, os autores compararam os métodos de SVM e Redes Neurais, e os resultados mostraram desempenho superior do SVM.

A pesquisa desenvolvida por Min *et al*. (2006) aponta a crescente aplicação do método SVM para o problema de análise financeira, como alternativa aos modelos de Regressão Logística e Análise Discriminante, mostrando resultados satisfatórios. O artigo propõe métodos para melhoria da performance do SVM por meio da utilização de Algoritmos Genéticos para otimização dos parâmetros do modelo.

Em sua publicação, Boyacioglu *et al*. (2009) propõem um método computacional para comparar os resultados obtidos na classificação de instituições financeiras por meio de métodos estatísticos multivariados e métodos relacionados a Redes Neurais. Foram selecionados 20 índices financeiros como variáveis preditoras no estudo. Quatro diferentes bancos de dados foram divididos em grupos de treinamento e validação dos modelos. Os métodos *Multi-layer Perceptron* e *Learning Vector Quantization* foram considerados como os mais adequados para prever a falência de bancos.

Fontalvo *et al.* (2012) estudaram a aplicação da Análise Discriminante para analisar a melhoria de indicadores financeiros no setor alimentício de Barranquilla. Inicialmente, foram descritos os tipos de índices utilizados e o modelo foi aplicado para distinguir o desempenho financeiro destas empresas nos anos de 2004 a 2009. Por meio da função discriminante, as variáveis de *Return on Assets*, Giro de Ativos e Endividamento Geral mostraram significativa diferença entre os períodos analisados.

Oh (2014) desenvolveu um método para identificação de variáveis para prever dificuldades financeiras de empresas de energia eólica na Coréia do Sul, comparando a acurácia dos métodos de Análise Discriminante e Regressão Logística. A base de dados utilizada no estudo consiste em 15 Companhias de energia eólica coreanas com ações listadas na Bolsa KOSDAQ, que tiveram baixa performance reportada em seus resultados de 2012. O autor identificou cinco índices financeiros estatisticamente significantes e superioridade dos resultados obtidos pela Regressão Logística em relação à Análise Discriminante.

Por fim, Tserng *et al.* (2014) elaboraram em estudo sobre a aplicação de em modelo de Regressão Logística para predição de insolvência de companhias de construção civil. Para analisar o desempenho do modelo, os autores calcularam a área abaixo da Curva ROC e concluíram que os resultados evidenciaram a robustez dos índices financeiros em discriminar as empresas, de forma satisfatória. O índice de Endividamento mostrou-se um importante indicador para a predição de inadimplência no setor de Construção Civil.

O presente trabalho visa explorar uma lacuna existente na literatura, ao utilizar métodos multivariados aplicados à análise de rentabilidade, dando relevante contribuição às publicações existentes, cujo foco consiste na análise de insolvência.

3 MATERIAL E MÉTODOS

Esta seção descreve a classificação da pesquisa, os materiais utilizados, a forma de coleta dos dados e sua utilização na análise multivariada.

3.1 CLASSIFICAÇÃO DA PESQUISA

Esta pesquisa visa aplicar métodos para predição do nível de rentabilidade de companhias automotivas, por meio da utilização de índices financeiros. Portanto, segundo Gil (2002), este trabalho se caracteriza como pesquisa aplicada, quanto ao propósito da investigação.

O método científico, de acordo com o objetivo do trabalho, é classificado como explicativo, pois tem a finalidade de identificar fatores que determinam ou que contribuem para os resultados das empresas em estudo. Gil (2002) pondera que este é o tipo de pesquisa que aprofunda o conhecimento da realidade, pois explica a razão de algum acontecimento.

O procedimento técnico a ser aplicado quando da execução do trabalho é a pesquisa bibliográfica, visto que esta pesquisa implica leituras sobre o assunto, auxiliando o entendimento sobre o problema e utilização de referências teóricas para execução da pesquisa.

No que se refere à abordagem do problema, a pesquisa pode ser classificada como quantitativa, pois conforme Kirk e Miller (1986), o pesquisador define claramente suas hipóteses e variáveis, usando-as para obter uma medição precisa dos resultados quantificáveis obtidos.

3.2 ENQUADRAMENTO DA PESQUISA

Com base nas definições expostas na seção anterior, a presente pesquisa foi enquadrada na estrutura disposta na FIGURA 11.

FIGURA 11 - ENQUADRAMENTO DA PESQUISA

FONTE: O autor (2016).

3.3 DELIMITAÇÃO DA PESQUISA

Neste trabalho, serão analisadas informações financeiras de montadoras de veículos com ações listadas em Bolsa de Valores. Com base em relatórios divulgados por sistemas de informações financeiras à disposição de investidores e demais *stakeholders*, torna-se possível a construção de um banco de dados com informações financeiras destas companhias.

De tal modo, foram selecionadas todas as companhias com capital aberto, as quais disponibilizam suas demonstrações contábeis, tais como o Balanço Patrimonial e a Demonstração do Resultado do Exercício, no sistema de informações financeiras utilizado neste trabalho.

A TABELA 4 apresenta a relação das 30 montadoras selecionadas neste estudo, as quais terão seu processo de coleta de dados explorado em detalhes na seção subsequente.

TABELA 4 - RELAÇÃO DE MONTADORAS SELECIONADAS

PAÍS	MONTADORAS	QUANTIDADE
ALEMANHA	BMW, DAIMLER, VOLKSWAGEN	3
CHINA	BAIC, BRILLIANCE, BYD, CHANGAN, DONGFENG, FAW, GEELY, GREATWALL, SAIC	9
CORÉIA DO SUL	HYUNDAI	1
ESTADOS UNIDOS	FORD, GM	2
FRANÇA	PEUGEOT, RENAULT	2
ÍNDIA	MAHINDRA, TATA	2
ITÁLIA	FIAT	1
JAPÃO	HONDA, ISUZU, FUJI, MAZDA, MITSUBISH, NISSAN, SUZUKI, TOYOTA	8
RÚSSIA	AVTOVAZ	1
SUÉCIA	VOLVO	1
TOTAL	**-**	**30**

FONTE: O autor (2016).

Conforme observado na TABELA 4, a análise financeira realizada neste trabalho, por meio de índices financeiros, engloba empresas sediadas em diferentes países. Tendo em vista que os índices são medidas adimensionais, por calcularem a razão entre contas contábeis das empresas, não há inconvenientes em se comparar companhias que consolidam seus resultados em diferentes moedas locais.

3.4 COLETA DE DADOS

De forma a alcançar os objetivos delineados nesta pesquisa, torna-se necessário o levantamento dos dados financeiros das montadoras citadas na seção anterior, os quais possuem restrições quanto à sua disponibilização.

Os métodos para análise de rentabilidade de empresas automotivas, aplicados neste trabalho consistem em técnicas estatísticas multivariadas, fundamentadas na utilização de índices financeiros extraídos do Balanço Patrimonial

e da Demonstração de Resultado do Exercício (DRE) de cada uma das companhias em estudo.

Ademais, Silva (2008) afirma que a avaliação por meio de índices financeiros é adequada para análises históricas e também para análises comparativas entre organizações. O autor pondera, ainda, a importância da utilização destas métricas para auxiliar a análise financeira de empresas.

As informações necessárias para a realização desta pesquisa foram coletadas por meio do sistema *on line MorningStar®*, disponível no endereço eletrônico *www.morningstar.com*. O sistema é utilizado por especialistas e investidores para avaliação de empresas, ações e fundos de investimento.

O sistema disponibiliza relatórios financeiros de companhias com ações listadas em Bolsa de Valores, possibilitando a coleta de informações sem a dependência de outros sistemas custosos ou restritos, tais como *Bloomberg* ou *Economática*. A FIGURA 12 mostra as variáveis coletadas no sistema *MorningStar®*.

FIGURA 12 - ÍNDICES FINANCEIROS COLETADOS

FONTE: O autor (2016).

Na literatura acadêmica são apresentados diversos indicadores econômico-financeiros que auxiliam na análise financeira, e alguns desses indicadores são apresentados por vários autores. Destarte, Matarazzo (2010) afirma que a definição do conjunto de índices para análise financeira possui diferentes abordagens, a depender dos objetivos de cada trabalho.

O critério para escolha dos índices financeiros a serem utilizados neste trabalho seguiu o direcionamento dos autores Assaf Neto (2007), Brealey *et al.*

(2001), Marion (2007), Matarazzo (2010) e Van Horne e Wachowicz (2008), no que se refere aos índices mais utilizados na análise financeira. Optou-se por coletar os índices financeiros abordados por estes autores e apresentados na seção 2.4, e que possuam informações disponíveis no sistema *MorningStar*®.

Optou-se pela inclusão da variável "Tamanho" no rol de indicadores financeiros a serem analisados, de forma a averiguar se a dimensão das empresas é uma variável capaz de influenciar sua rentabilidade, tal como mencionado por Matarazzo (2010) e Silva (2008). Esta variável é medida como o *logaritmo neperiano* do faturamento anual da empresa.

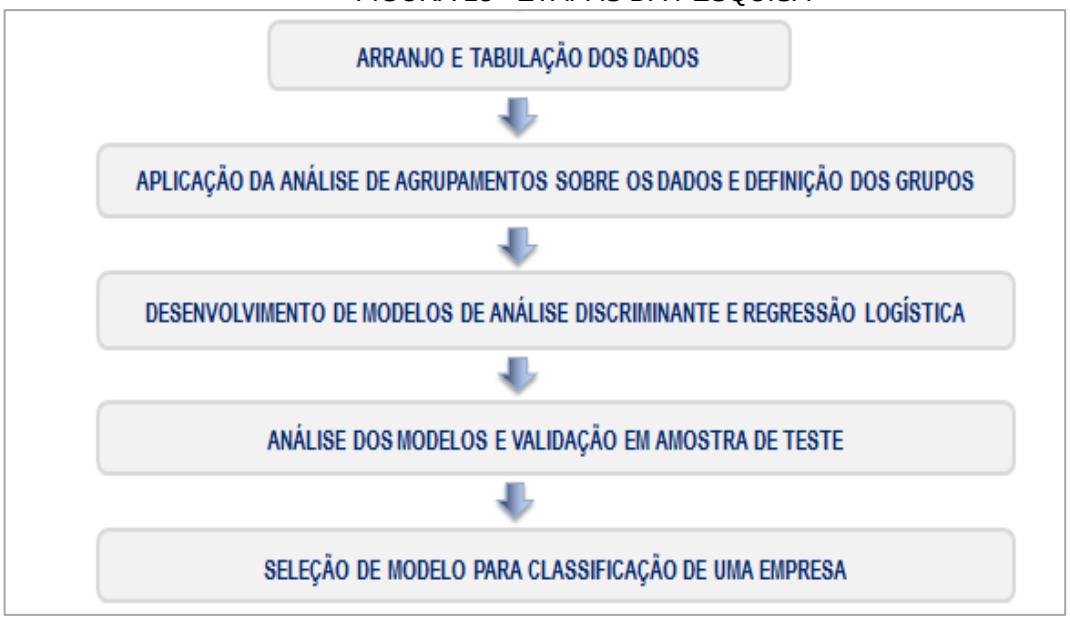
Ademais, o período de avaliação utilizado neste trabalho engloba os exercícios contábeis dos últimos 10 anos das trinta montadoras selecionadas, totalizando 300 observações com as informações financeiras de cada montadora para cada ano de avaliação.

3.5 SISTEMATIZAÇÃO E ANÁLISE DOS DADOS

Para aplicação da análise multivariada sobre a rentabilidade das empresas automotivas, serão utilizadas as 18 variáveis expostas na FIGURA 12.

De tal modo, a FIGURA 13 elenca os procedimentos metodológicos desempenhados ao longo deste trabalho, após a coleta dos dados necessários.

FIGURA 13 - ETAPAS DA PESQUISA



FONTE: O autor (2016).

3.5.1 Arranjo e tabulação dos dados

Inicialmente, arquivos com os índices financeiros de cada uma das 30 companhias selecionadas, referentes aos 10 anos de avaliação, foram extraídos do sistema *MorningStar®,* em formato '*csv'* e armazenados individualmente.

Posteriormente, todas as informações existentes nestes arquivos foram tabuladas em formato de planilhas eletrônicas por meio do aplicativo *Microsoft Excel* e agrupadas em um único arquivo (banco de dados), totalizando:

- 300 registros (linhas), representando dez anos de observação para cada uma das trinta montadoras selecionadas;
- 18 variáveis de *input* (colunas) para os modelos multivariados, representando os 18 índices financeiros descritos na FIGURA 12.

Com o intuito de certificar a integridade das informações, lançou-se mão de uma análise de possíveis registros em branco (*missing*) no banco de dados, de forma a não causar qualquer distorção ou interferência nas estimativas realizadas pelos modelos. Após esta análise, a quantidade de observações do banco de dados foi reduzida de 300 para 292 registros.

Então, a matriz de dados foi segregada em bases de dados distintas:

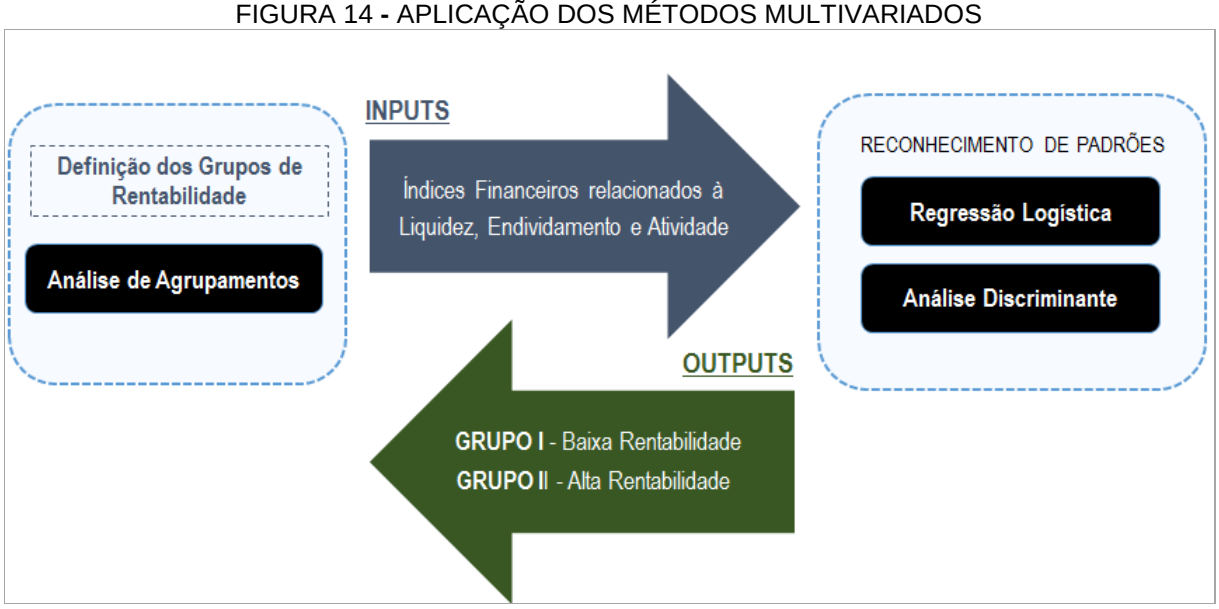
- Base de dados *"Clustering"*: matriz com os 292 registros, e 3 colunas (variáveis) com os índices de rentabilidade ROE, ROA e ROI para cada empresa;
- Base de dados *"*Reconhecimento de Padrões*"*: matriz com os 292 registros, e 15 colunas (variáveis) com os demais índices financeiros de cada empresa.

Por fim, as variáveis de cada base de dados foram validadas, a fim de identificar possíveis anomalias ou contradições nos registros, e não foram identificadas incoerências nas informações. Ademais, a análise de possíveis *outliers* não apontou registros fora dos padrões das variáveis analisadas.

3.6 APLICAÇÃO DE MÉTODOS ESTATÍSTICOS MULTIVARIADOS

As bases de dados "*Clustering*" e "Reconhecimento de Padrões", devidamente homologadas pelos processos de validação mencionados anteriormente, são submetidas a análises estatísticas. A FIGURA 14 elucida o funcionamento dos métodos multivariados, no âmbito desta pesquisa.

FIGURA 14 - APLICAÇÃO DOS MÉTODOS MULTIVARIADOS



FONTE: O autor (2016).

A modelagem estatística desta pesquisa foi realizada no *software R, versão 3.2.3* (R CORE TEAM, 2015), e encontra-se estruturada em linha com os procedimentos descritos nas seções a seguir.

3.6.1 Definição dos grupos de rentabilidade

Em linhas gerais, os métodos de reconhecimento de padrões aplicados neste trabalho visam alocar montadoras de automóveis em dois grupos possíveis: baixa ou alta rentabilidade. Assim, primeiramente, definem-se os níveis de rentabilidade que indicam se o desempenho de uma montadora deve ser classificado como bom ou ruim.

Não foi identificado na literatura acadêmica, um consenso acerca da forma de se definir uma fronteira entre alta e baixa rentabilidade de empresas, portanto, optou-se pela utilização da técnica estatística de Análise de Agrupamentos para

definir níveis de rentabilidade. Aplica-se este método de forma a segregar as empresas em dois grupos distintos:

- Baixa Rentabilidade - *Grupo I*;
- Alta Rentabilidade - *Grupo II*.

De tal modo, os dois grupos de empresas são definidos pela aplicação da técnica de Análise de Agrupamentos sobre a base de dados *"Clustering",* definida na seção 3.5.1, ou seja, são utilizados os índices financeiros ROA, ROE e ROI para separar as empresas de alta ou baixa rentabilidade.

O critério para escolha destas métricas consiste no fato de que, dentre os índices de rentabilidade apontados pelo referencial teórico, estes coeficientes dispõem de informações no sistema *MorningStar*® e permitem a análise dos lucros auferidos pelas empresas após dedução de despesas operacionais e financeiras.

De tal modo, apesar da possível correlação existente entre estes índices de rentabilidade, pretende-se incorporar ao procedimento de *clustering* a diversidade de informações necessárias para a distinção dos grupos de forma satisfatória.

A distância euclidiana será a medida de similaridade adotada e diferentes tipos de algoritmos aglomerativos são testados com a finalidade de construir os dois agrupamentos, conforme métodos apresentados na seção 2.5.3. Definiu-se a quantidade de grupos *a priori*, em linha com os objetivos definidos nesta pesquisa.

Em seguida, é aplicada a análise de variância multivariada para avaliação da formação de dois grupos distintos, com base no valor do Lambda de Wilks.

O método de ligação escolhido para agrupamento será aquele que originar melhor distinção entre os grupos e menor variância dentro de cada *cluster*.

3.6.2 Predição de rentabilidade: Análise Discriminante de Fisher

Após os resultados gerados pela Análise de Agrupamentos, a variável "Grupo*"* é adicionada na base de dados *"*Reconhecimento de Padrões*"* definida na seção 3.5.1, indicando se a empresa possui alta ou baixa rentabilidade.

A Análise Discriminante indicará os índices financeiros que discriminam ou explicam em qual grupo a empresa deve ser alocada. Em linhas gerais, busca-se

identificar quais variáveis financeiras discriminam se uma empresa possui baixa rentabilidade (Grupo I) ou alta rentabilidade (Grupo II).

Com base na função discriminante linear de Fisher, é possível encontrar *scores* para cada uma das variáveis explicativas do modelo e verificar a contribuição de cada uma para a alocação da empresa em um dos dois grupos.

De tal modo, serão analisados os pressupostos necessários para a aplicação da função discriminante linear de Fisher. O pressuposto de igualdade das matrizes de covariância dos grupos será avaliado por meio de gráficos *Box-Plot* e validado com a aplicação do teste de igualdade de variâncias de Bartlett.

O poder preditivo do modelo será analisado pelo percentual de acertos observados na Matriz de Confusão e a área sob a curva ROC, ao se classificar as empresas da amostra de teste. Em seguida será executada a análise de variância multivariada, em relação ao valor do Lambda de Wilks, para verificar se os grupos podem ser separados de forma satisfatória por suas variáveis discriminantes.

3.6.3 Predição de rentabilidade: Regressão Logística

A capacidade preditiva do método de Regressão Logística aplicado neste trabalho é comparada aos resultados obtidos pela Análise Discriminante. O modelo logístico também é aplicado sobre a base de dados “Reconhecimento de Padrões*”* e as empresas são classificadas em um dos grupos definidos na Análise de Agrupamentos, a partir de variáveis explicativas selecionadas pelo modelo.

O procedimento *stepwise* para seleção de variáveis visa incorporar ao modelo somente as variáveis que possuam alto poder para definir se uma determinada empresa possui alta ou baixa rentabilidade. As estimativas dos coeficientes da Regressão Logística, referentes às variáveis explicativas selecionadas, são calculadas pelo método da Máxima Verossimilhança e avaliadas conforme resultados dos testes de Razão de Verossimilhança e Wald.

O poder preditivo do modelo será analisado com base no percentual de acertos observados na Matriz de Confusão e na área sob a curva ROC, ao se classificar as empresas da amostra de testes. A acurácia do modelo também será avaliada pelas estatísticas Pseudo R^2 de McFadden, R^2 de Cox e Snell e R^2 de Nagelkerke. Por fim, será realizada a análise de resíduos.

3.6.4 Seleção do modelo para classificação de empresas

Após comparação entre os resultados obtidos pelos métodos de Regressão Logística e Análise Discriminante, será selecionado o método que apresente maior poder preditivo ao discriminar montadoras de veículos com “alta rentabilidade” e “baixa rentabilidade”. Tal escolha está atrelada aos resultados obtidos na classificação da amostra de testes, no que se refere às métricas de avaliação apresentadas na seção 2.6.3, ou seja:

- Percentual de acertos ao classificar empresas;
- Área sob a curva ROC.

4 RESULTADOS E DISCUSSÕES

No decorrer deste capítulo, serão empregados os procedimentos necessários ao cumprimento dos objetivos delineados nesta pesquisa. Neste sentido, serão discutidos os resultados obtidos e os aspectos técnicos à luz do arcabouço teórico apresentado ao longo deste trabalho.

Primeiramente, será descrita a utilização do método de Análise de Agrupamentos para definição de *clusters*. Posteriormente, as técnicas de Análise Discriminante de Fisher e Regressão Logística são aplicadas para reconhecimento de padrões na rentabilidade das montadoras multinacionais.

4.1 FORMAÇÃO DE GRUPOS DE RENTABILIDADE

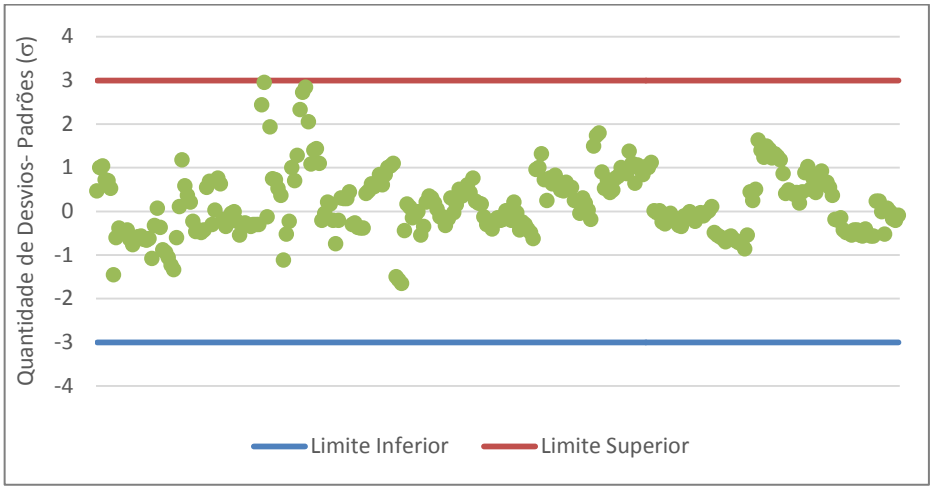
Conforme abordado nas seções anteriores, a Análise de Agrupamentos busca agrupar elementos baseando-se na similaridade existente entre os mesmos. Os agrupamentos são determinados de forma a obter-se homogeneidade dentro dos grupos e diferença significativa entre os *clusters*.

Neste contexto, com o objetivo de segregar cada montadora em dois grupos distintos (alta rentabilidade e baixa rentabilidade), optou-se pela estruturação de uma Análise de Agrupamentos por meio de métodos hierárquicos, com base nos indicadores *Return on Assets*, *Return on Equity* e *Return On Investment*.

Após a seleção da base de dados, procedeu-se à análise de possíveis *outliers* que pudessem, eventualmente, implicar em inconsistências quando da formação dos grupos.

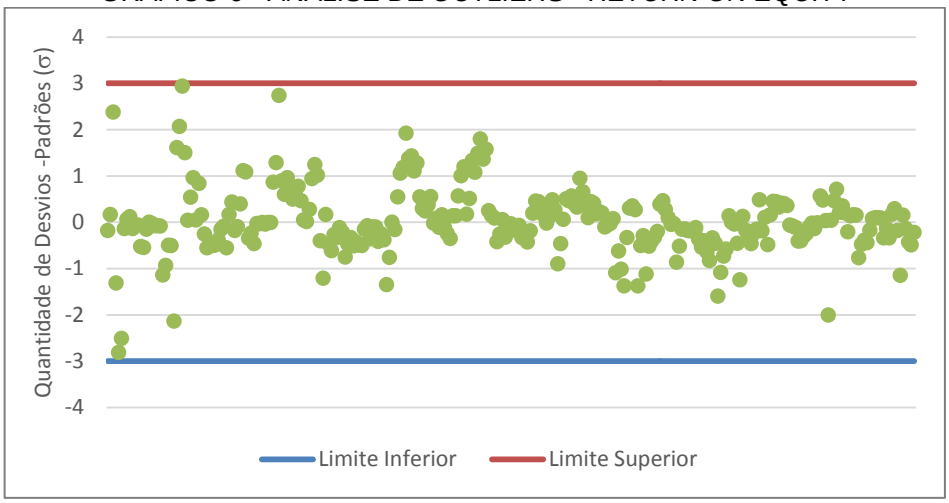
Os GRÁFICOS 5, 6 e 7 apresentam análises gráficas de *outliers* a partir das observações padronizadas para cada uma das variáveis utilizadas no processo de *clustering*. Foram definidos intervalos superiores e inferiores com base em uma distribuição normal padronizada.

GRÁFICO 5 - ANÁLISE DE *OUTLIERS* - *RETURN ON ASSETS*



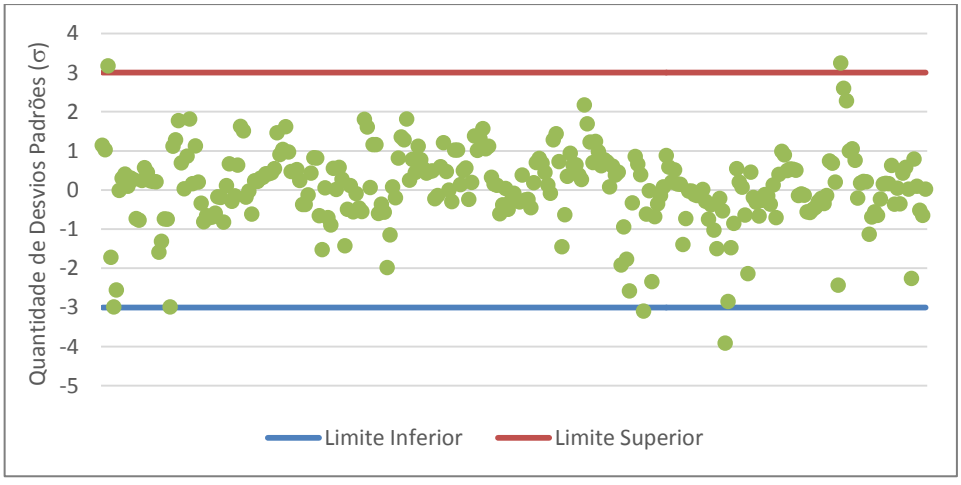
FONTE: O autor (2016).

GRÁFICO 6 - ANÁLISE DE *OUTLIERS* - *RETURN ON EQUITY*



FONTE: O autor (2016).

GRÁFICO 7 - ANÁLISE DE *OUTLIERS* - *RETURN ON INVESTMENT*



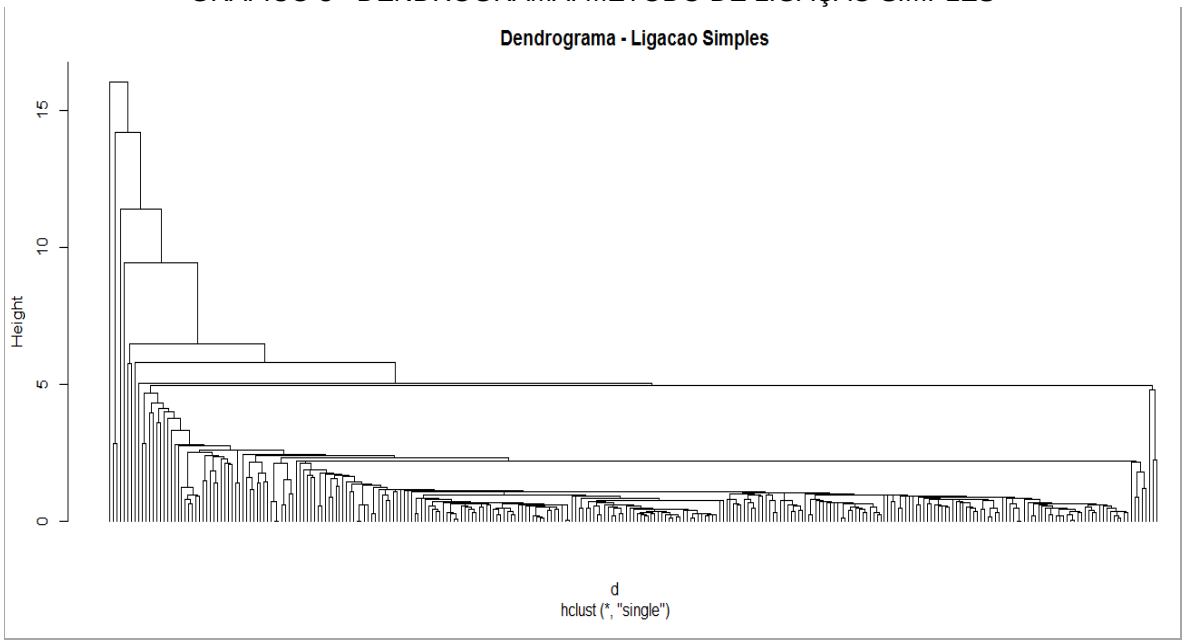
FONTE: O autor (2016).

Os resultados denotam a existência de alguns pontos além dos limites previamente definidos, os quais são interpretados como *outliers*. Após análise destas observações, constatou-se que tais registros não devem ser excluídos.

Em seguida, a Análise de Agrupamentos foi estruturada no *software* estatístico *software R, versão 3.2.3* (R CORE TEAM, 2015), e cada algoritmo hierárquico foi avaliado a partir de dendrogramas e da Análise de Variância Multivariada (MANOVA). Adotou-se a distância euclidiana como medida de similaridade, por apresentar melhores resultados em relação às outras medidas.

No que tange ao método de Ligação Simples, o GRÁFICO 8 denota a inviabilidade de se formar dois grupos distintos, em virtude da existência de longas cadeias. Há um primeiro grupo de um ou mais elementos que passa a incorporar, a cada iteração, um grupo de apenas um elemento, tornando difícil a determinação de um nível de corte para classificar as observações.

GRÁFICO 8 - DENDROGRAMA: MÉTODO DE LIGAÇÃO SIMPLES



FONTE: O autor (2016).

Na TABELA 5, a MANOVA ratifica a impossibilidade em se formar dois *clusters* estatisticamente distintos. O *p-valor* não indica a rejeição da hipótese nula de que os grupos são estatisticamente iguais ao nível de significância de 0,1%. O valor do Lambda de Wilks reflete a alta variância existente dentro dos grupos, a qual indica formação de *clusters* não homogêneos.

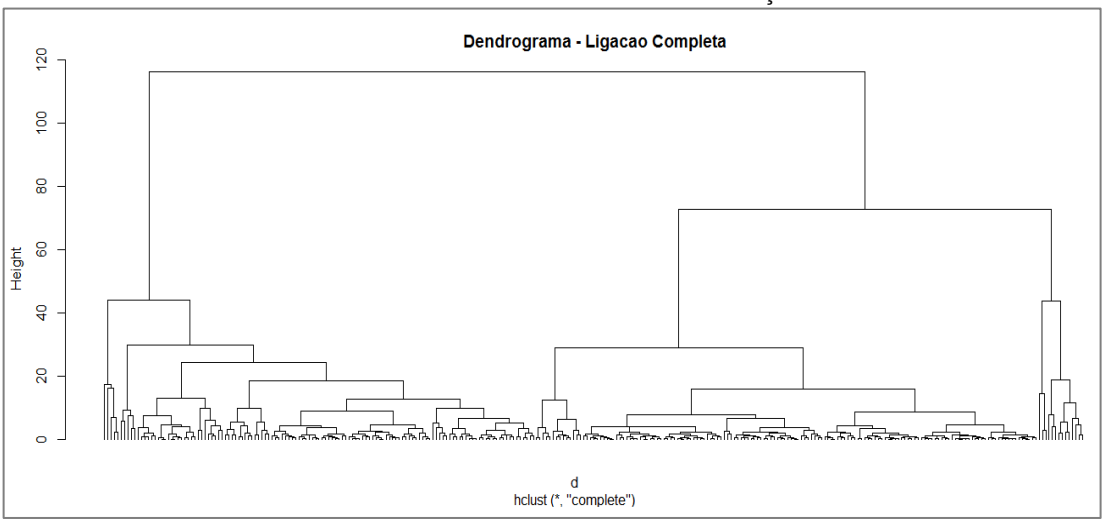
TABELA 5 - MANOVA: MÉTODO DE LIGAÇÃO SIMPLES

Lambda de Wilks	F	*p-valor*
0,9605	3,96	0,009

FONTE: O autor (2016).

O método de Ligação Completa obteve um desempenho superior no que se refere à formação dos dois grupos de empresas. No dendrograma apresentado no GRÁFICO 9 é possível notar a formação de dois grupos com número de elementos melhor definidos e com menor formação de encadeamentos.

GRÁFICO 9 - DENDROGRAMA: MÉTODO DE LIGAÇÃO COMPLETA



FONTE: O autor (2016).

No que se refere à Análise de Variância, os dados da TABELA 6 mostram o desempenho satisfatório na formação dos *clusters*. O *p-valor* indica a rejeição da hipótese nula de que os grupos não possuem diferença significativa. O valor do Lambda de Wilks sugere a formação de grupos homogêneos (baixa variância intragrupos), enquanto a estatística *F* indica a formação de dois grupos distintos.

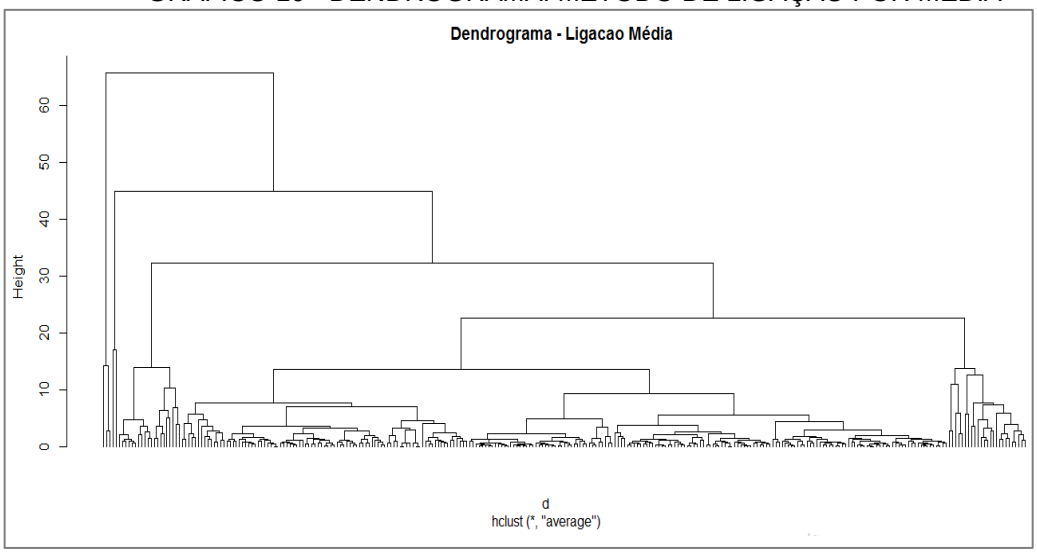
TABELA 6 - MANOVA: MÉTODO DE LIGAÇÃO COMPLETA

Lambda de Wilks	F	*p-valor*
0,4946	98,44	< 0,0001

FONTE: O autor (2016).

O dendrograma do método de Ligação por Média, exposto no GRÁFICO 10, não apresentou resultados satisfatórios. Tal como na Ligação Simples, nota-se a existência de encadeamentos, ou seja, em ambos os algoritmos há companhias com índices de rentabilidade distintos classificadas em um mesmo grupo, impossibilitando a separação entre empresas com alta e baixa rentabilidade.

GRÁFICO 10 - DENDROGRAMA: MÉTODO DE LIGAÇÃO POR MÉDIA



FONTE: O autor (2016).

Ademais, este método obteve resultados superiores à Ligação Simples, porém, em nível inferior ao método de Ligação Completa. O *p-valor* indica a rejeição da hipótese nula de que os grupos não possuem diferenças significativas e a estatística *F* sugere a formação de dois grupos distintos. Contudo, o valor do Lambda de Wilks ainda reflete a existência de alta variância dentro dos grupos, como pode ser evidenciado na TABELA 7.

TABELA 7 - MANOVA: MÉTODO DE LIGAÇÃO MÉDIA

Lambda de Wilks	**F**	***p-valor***
0,7820	26,86	< 0,0001

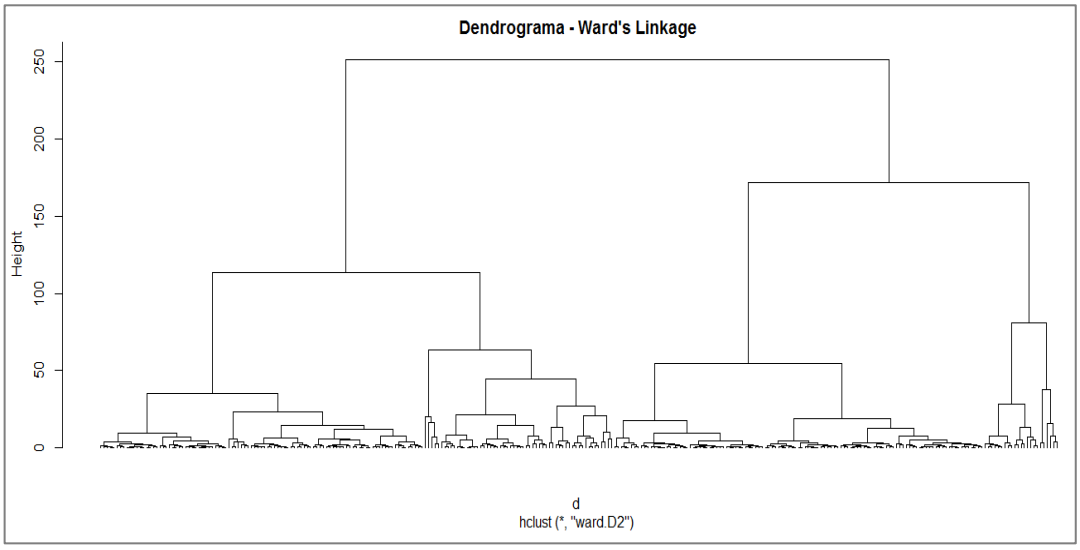
FONTE: O autor (2016).

Por fim, o Método de Ward obteve resultados superiores aos demais algoritmos. No dendrograma do GRÁFICO 11 é possível notar a ausência de

encadeamentos e formação de grupos compactos, com maior clareza quanto à quantidade de elementos presentes em cada grupo.

Neste caso, as junções iniciais entre os *clusters* são pequenas, o que indica distâncias próximas de zero. Conforme o número de iterações aumenta, estes *clusters* são absorvidos por grupos maiores, com distâncias mais elevadas, gerando junções em pontos mais altos. Portanto, nota-se a viabilidade em se formar dois *clusters* a partir da similaridade existente entre as empresas.

GRÁFICO 11 - DENDROGRAMA: MÉTODO DE WARD



FONTE: O autor (2016).

Os resultados alcançados na Análise de Variância ratificam a superioridade do Método de Ward no procedimento de *clustering* ora apresentado, uma vez que se obtém forte distinção entre os grupos, com alto valor para a estatística *F* e baixa variância dentro de cada *cluster,* representada pelo Lambda de Wilks. Estes dados podem ser visualizados na TABELA 8.

TABELA 8 - MANOVA: MÉTODO DE WARD

Lambda de Wilks	**F**	***p-valor***
0,4933	98,93	< 0,0001

FONTE: O autor (2016).

O método de Ward pode resultar em agrupamentos de tamanhos aproximadamente iguais devido à minimização de variação interna. Em cada

estágio, combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos (BUSSAB *et al*., 1999).

A TABELA 9 resume os resultados de todos os métodos avaliados.

TABELA 9 - COMPARATIVO ENTRE MÉTODOS HIERÁRQUICOS

Método	Lambda de Wilks	F	*p-valor*
Ligação Simples	0,9605	3,96	0,009
Ligação Completa	0,4946	98,44	<0,0001
Ligação por Média	0,7812	26,86	<0,0001
Método de Ward	**0,4933**	**98,94**	**<0,0001**

FONTE: O autor (2016).

O conhecimento do pesquisador sobre o assunto possui grande importância para a interpretação conceitual dos *clusters*, para fundamentação dos dados pertencentes a cada agrupamento encontrado. Ademais, o *clustering* sempre encontrará grupos em um conjunto de dados, mesmo que não exista uma descrição relacionada às classes definidas no escopo da pesquisa (HAIR *et al.*, 2010).

Neste contexto, os resultados alcançados pelo Método de Ward e pela Ligação Completa geraram resultados satisfatórios. Optou-se pelo Método de Ward, pois a classificação gerada por este algoritmo, além de melhores resultados estatísticos, indicou maior coerência técnica na definição do limiar entre as montadoras com alta rentabilidade e aquelas com desempenho inferior.

A TABELA 10 apresenta os valores médios de cada variável para os grupos formados. O *Grupo I* consiste nas empresas com baixa rentabilidade, enquanto o *Grupo II* é formado por empresas com desempenho superior.

TABELA 10 - ÍNDICES DE RENTABILIDADE: MÉTODO DE WARD

	Grupo I	Grupo II
ROE - Return on Equity	1,5%	6,7%
ROA - Return On Assets	0,4%	1,1%
ROI - Return on Investment	4,6%	19,0%

FONTE: O autor (2016).

A Análise de Agrupamentos apresentada nesta seção permitiu a construção de uma abordagem robusta para definição de níveis de rentabilidade das montadoras de veículos, dispensando a utilização de critérios subjetivos.

A formação destes *clusters* permite que uma empresa seja classificada frente ao desempenho financeiro de seus concorrentes. Assim, uma empresa será considerada no *Grupo II* caso os retornos gerados por suas atividades estejam em um patamar elevado, se comparado às demais montadoras.

4.2 RECONHECIMENTO DE PADRÕES

Nesta seção serão apresentadas a aplicação e a validação de uma metodologia capaz de prever se determinada empresa deve ser incluída no grupo de companhias com rentabilidade superior, definido na seção anterior pela Análise de Agrupamentos.

De tal modo, será utilizado o *software* estatístico *software R, versão 3.2.3* (R CORE TEAM, 2015), para modelar os métodos multivariados de Análise Discriminante e Regressão Logística no reconhecimento de padrões e classificação das montadoras de veículos em estudo. A partir de um conjunto de índices financeiros, espera-se identificar quais destas variáveis são determinantes para distinguir as montadoras com rentabilidade superior daquelas com baixo desempenho, além de prever seu nível de rentabilidade.

Optou-se pela segregação do banco de dados em um conjunto de treinamento e outro de teste, permitindo uma visão mais realista do erro a ser cometido pelo método de classificação empregado (HAIR *et al.*, 2010). Os parâmetros dos modelos foram estimados em um ambiente de treinamento e, posteriormente, a amostra de teste é selecionada para averiguação da qualidade do ajustamento dos modelos.

No que se refere à composição dos registros do banco de dados em amostras de treinamento e de teste, existe flexibilidade quanto à definição dos percentuais, havendo a opção de certos autores por diferentes composições (HAIR *et al.*, 2010). Neste trabalho, optou-se pela seguinte composição:

- Amostra de Treinamento: 60% (174 registros);
- Amostra de Teste: 40% (118 registros).

De forma similar aos modelos tradicionais de *creditscore*, a variável resposta (dependente) dos modelos para reconhecimento de padrões apresentados neste trabalho consiste em uma variável dicotômica, denominada “Grupo Predito”, que pode assumir os valores 0 ou 1, conforme abaixo:

- “Grupo Predito” = 0, caso a rentabilidade da empresa seja predita como inferior, ou seja, a observação é alocada no *Grupo I*, previamente definido pela Análise de Agrupamentos;
- “Grupo Predito” = 1, caso a rentabilidade da empresa seja predita como superior, ou seja, a observação é alocada no *Grupo II*, previamente definido pela Análise de Agrupamentos.

Para classificar as observações conforme seu nível de rentabilidade seleciona-se as variáveis explicativas que possam influenciar a situação financeira das montadoras multinacionais em estudo. As variáveis inicialmente testadas são os índices financeiros da base de dados “Reconhecimento de Padrões”, definidos na seção 3.4.

4.2.1 Análise Discriminante de Fisher

A aplicação do método de Análise Discriminante de Fisher neste trabalho tem ênfase na derivação de uma regra que pode ser utilizada para classificação de montadoras multinacionais. Ademais, torna-se possível o conhecimento prévio das variáveis que influenciam o nível de rentabilidade destas companhias.

4.2.1.1 Seleção de variáveis

Inicialmente, os índices financeiros são pré-selecionados como possíveis variáveis preditoras da função discriminante linear de Fisher. O critério utilizado para seleção de variáveis consiste na avaliação do Lambda de Wilks, sendo incluídos no modelo somente aqueles índices financeiros com poder discriminatório, descartando-se as demais variáveis.

Dentre as 15 variáveis pré-selecionadas no ambiente de treinamento, foram selecionados 10 índices financeiros, os quais possuem significância estatística, com

base no valor do Lambda de Wilks. A TABELA 11 mostra os resultados obtidos pelo procedimento de seleção de variáveis.

TABELA 11 - SELEÇÃO DE VARIÁVEIS: ANÁLISE DISCRIMINANTE

Variável	Lambda de Wilks	F	*p-valor*
Margem Operacional	0,6874	78,20	<0,0001
Giro do Ativo Total	0,6419	47,69	<0,0001
Endividamento de Longo Prazo	0,6335	32,77	<0,0001
Imobilização sobre Patrimônio Líquido	0,6195	25,94	<0,0001
Liquidez Corrente	0,6051	21,92	<0,0001
Liquidez Imediata	0,5715	20,86	<0,0001
Giro de Estoque	0,5438	19,89	<0,0001
Tamanho	0,5070	20,05	<0,0001
Giro do Ativo Imobilizado	0,4813	19,63	<0,0001
Prazo Médio de Pagamento de Compras	0,4636	18,86	<0,0001

FONTE: O autor (2016).

4.2.1.2 Pressupostos da Análise Discriminante

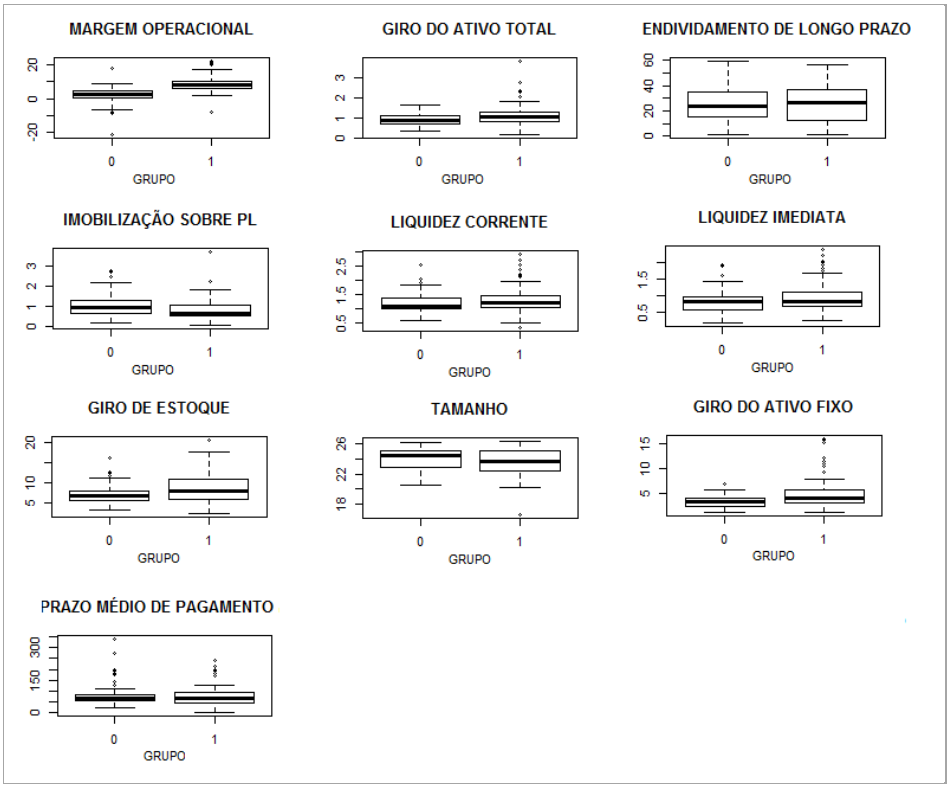
A fim de certificar a validade da regra de classificação desenvolvida, foram analisados os pressupostos do modelo de Análise Discriminante. Segundo Johnson e Wichern (2007), a utilização da função discriminante linear de Fisher dispensa o pressuposto de normalidade multivariada dos dados, restando o cumprimento dos seguintes pressupostos:

1) Homogeneidade das matrizes de covariância;
2) Ausência de problemas relacionados à multicolinearidade.

Primeiramente, cada uma das variáveis explicativas selecionadas foi analisada em relação ao pressuposto da “Homogeneidade das Matrizes de

Covariância". Desta forma, foi realizada a análise de gráficos *Box-Plot* para cada uma das variáveis independentes, conforme FIGURA 15.

FIGURA 15 - BOX-PLOT PARA ANÁLISE DE VARIÂNCIAS



FONTE: O autor (2016).

Os gráficos apresentados na FIGURA 15 denotam homogeneidade de variâncias para os dois grupos de empresas, pois de forma geral, não se observou diferenças significativas entre as variâncias dos conjuntos à esquerda (baixa rentabilidade) e à direita (alta rentabilidade).

De forma a ratificar a existência de homogeneidade de variâncias, optou-se pela realização do *Teste de Bartlett* ao nível de significância de 5%. As seguintes hipóteses foram testadas:

- H_0: Não há diferença significativa entre a variância dos grupos;
- H_1: Há diferença significativa entre a variância dos grupos.

A TABELA 12 apresenta os resultados obtidos no teste de Bartlett.

TABELA 12 - TESTE DE HOMOGENEIDADE DE VARIÂNCIAS

Variável	*p-valor*
Margem Operacional	0,2607
Giro do Ativo Total	0,2560
Endividamento de Longo Prazo	0,6505
Imobilização sobre Patrimônio Líquido	0,5583
Liquidez Corrente	0,1812
Liquidez Imediata	0,1684
Giro de Estoque	0,1227
Tamanho	0,2942
Giro do Ativo Imobilizado	0,1713
Prazo Médio de Pagamento de Compras	0,1185

FONTE: O autor (2016).

Os resultados apresentados na TABELA 12 foram satisfatórios e observou-se homogeneidade de variância entre os dois grupos para todas as variáveis, em linha com a indicação dos gráficos *Box-Plot*. Nenhuma das variáveis preditoras apresentou *p-valor* menor que o nível de significância de 0,05, isto é, não se rejeita a hipótese nula H_0.

Portanto, considera-se cumprimento do pressuposto de “Homogeneidade das Matrizes de Covariância*”* da função discriminante linear de Fisher.

No que se refere ao pressuposto da “Ausência de Colinearidade”, o método de seleção de variáveis utilizado*,* ao incluir uma nova variável no modelo, exclui automaticamente aquelas que são altamente correlacionadas como índice selecionado. Portanto, inconsistências geradas pela adição de variáveis colineares são tratadas pelo modelo e tal pressuposto é atendido.

4.2.1.3 Função discriminante linear de Fisher

Após averiguação das premissas do modelo, deu-se prosseguimento ao desenvolvimento da função discriminante linear de Fisher. Para tal, utilizou-se a amostra de treinamento, composta por 174 observações selecionadas aleatoriamente do banco de dados.

Observou-se a existência de 80 registros pertencentes ao *Grupo I* (baixa rentabilidade) e 94 observações do *Grupo II* (alta rentabilidade). A probabilidade de classificação *a priori* para cada grupo foi calculada conforme TABELA 13.

TABELA 13 - PROBABILIDADES *A PRIORI* E FREQUÊNCIAS

	Grupo I	***Grupo II***
Probabilidades a priori	46,0%	54,0%
Quantidade de Registros	80	94

FONTE: O autor (2016).

Em seguida, foram estimados os coeficientes da função discriminante linear de Fisher, a partir dos quais se torna possível a classificação de novas empresas em um destes grupos existentes, com base nas variáveis explicativas, selecionadas pelo critério do Lambda de Wilks.

A TABELA 14 apresenta os coeficientes estimados do modelo:

TABELA 14 - FUNÇÃO DISCRIMINANTE LINEAR DE FISHER

Variável	**Coeficiente Estimado**
Margem Operacional	0,2311
Giro do Ativo Total	0,8133
Endividamento de Longo Prazo	0,0534
Imobilização sobre Patrimônio Líquido	-0,1889
Liquidez Corrente	-3,2910
Liquidez Imediata	2,2090
Giro de Estoque	0,0561
Tamanho	-0,2816
Giro do Ativo Imobilizado	0,1632
Prazo Médio de Pagamento de Compras	-0,0056

FONTE: O autor (2016).

4.2.1.4 Avaliação do desempenho do modelo

De forma a avaliar o desempenho do modelo, calculou-se o percentual de acertos ao se classificar as empresas da amostra de teste, a qual é composta por 40% dos registros do banco de dados original. A TABELA 15 apresenta a Matriz de Confusão, utilizada para cálculo dos percentuais de acertos.

TABELA 15 - MATRIZ DE CONFUSÃO: ANÁLISE DISCRIMINANTE

GRUPO	*I*	*II*
I	54	6
II	7	51

FONTE: O autor (2016).

O percentual geral de acertos do modelo em classificar as empresas da amostra de teste, alcançou a taxa de 89,0%. A TABELA 16 revela menor probabilidade de classificação incorreta para o *Grupo I*, ou seja, empresas com performance inferior.

TABELA 16 - TAXA DE ACERTOS: ANÁLISE DISCRIMINANTE

Grupo I	***Grupo II***
90,0%	87,9%

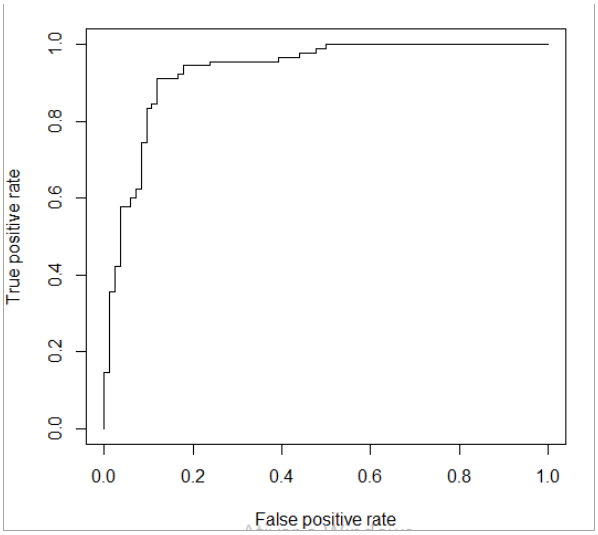
FONTE: O autor (2016).

A área abaixo da curva de ROC fornece uma medida de discriminação, a qual indica a probabilidade de as empresas que possuem melhor desempenho sejam classificadas no grupo superior àquelas de menor rentabilidade.

A área calculada sob a curva ROC alcançou o valor de 0,9406, ratificando o desempenho satisfatório da função discriminante linear. Segundo Hosmer e Lemeshow (2000), modelos com área sob a curva ROC superior a 0,90 possuem excelente poder discriminatório.

O GRÁFICO 12 apresenta a curva ROC gerada a partir da classificação da amostra de testes.

GRÁFICO 12 - CURVA ROC: ANÁLISE DISCRIMINANTE



FONTE: O autor (2016).

Em seguida, para ajuizar se os grupos foram devidamente separados, lançou-se mão da Análise de Variância Multivariada (MANOVA). Com base na TABELA 17, observa-se que os resultados são satisfatórios, uma vez que a hipótese nula de que há igualdade entre os grupos é rejeitada, com base nas métricas Lambda de Wilks, Traço de Pillai e T de Hotteling-Lawley.

TABELA 17 - MANOVA: ANÁLISE DISCRIMINANTE

Métrica	Estatística	*p-valor*
Lambda de Wilks	0,4972	<0,0001
Traço de Pillai	0,5028	<0,0001
Hotteling-Lawley	1,011	<0,0001

FONTE: O autor (2016).

4.2.2 Regressão logística

O método multivariado de Regressão Logística irá utilizar os mesmos dados empregados no método de Análise Discriminante, ou seja, as mesmas amostras de treinamento e de testes serão utilizadas para calibração e validação do modelo, respectivamente.

4.2.2.1 Seleção de variáveis

Tal como no desenvolvimento da função discriminante linear de Fisher, descrito na seção anterior, os índices financeiros são pré-selecionados como possíveis variáveis preditoras do modelo de Regressão Logística.

A escolha das variáveis será feita por intermédio do método *forward stepwise*, utilizado para seleção de variáveis no método de Regressão Logística, o qual permite que as variáveis sejam incluídas a cada passo, para otimizar o modelo e eliminar problemas relacionados à multicolinearidade (HAIR *et al.*, 2010).

Inicialmente foram selecionadas todas as 15 variáveis da amostra de treinamento. Após a execução do procedimento, o número de variáveis explicativas foi reduzido a um total de 9 índices financeiros, com significativo poder preditivo.

Quando o Método dos Mínimos Quadrados é utilizado estimar os parâmetros de um modelo com resultado dicotômico, os estimadores não apresentam as propriedades estatísticas desejáveis (HAIR *et al.*, 2010). Sendo assim, os parâmetros $\beta_0, \beta_1, \ldots, \beta_9$ foram ajustados pelo método de Máxima Verossimilhança, implementado de maneira iterativa para encontrar as estimativas dos coeficientes do modelo de Regressão Logística.

A TABELA 18 apresenta as variáveis selecionadas pelo procedimento *stepwise* e os coeficientes estimados para o modelo de Regressão Logística.

TABELA 18 - VARIÁVEIS SELECIONADAS PELO MÉTODO *STEPWISE*

Variável	Coeficientes Estimados ($\hat{\beta}$)
Constante	8,6135
Tamanho	-0,7504
Margem Operacional	0,9769
Liquidez Seca	9,5372
Participação de Capitais de Terceiros	0,1083
Liquidez Corrente	-11,7070
Prazo Médio de Pagamento de Compras	-0,0218
Giro do Ativo Imobilizado	0,8764
Giro do Ativo Total	3,7178
Composição do Endividamento	-0,0523

FONTE: O autor (2016).

4.2.2.2 Testes de significância

Após estimar os coeficientes de regressão, deve-se averiguar a significância das variáveis preditoras para dar prosseguimento às análises necessárias. Neste sentido, a informação sobre as variáveis independentes permite que sejam feitas previsões estatisticamente válidas (HOSMER; LEMESHOW, 2000).

Esta averiguação consiste na realização de testes de hipóteses para avaliação do grau de significância de cada variável. Portanto, procede-se ao seguinte teste de hipóteses:

- H_0: $\underline{\beta}_{(9x1)} = 0$
- H_1: $\underline{\beta}_{(9x1)} \neq 0$

O teste da Razão de Verossimilhança comparou os valores observados da variável resposta com os valores preditos obtidos nos modelos com e sem a variável em questão, conforme TABELA 19. *O* modelo *I* é formulado somente com o intercepto β_0, enquanto o modelo *II* possui o vetor $\underline{\beta}_{(9x1)}$ com os coeficientes estimados para as 9 variáveis selecionados no método *stepwise.*

TABELA 19 - TESTE DE RAZÃO DE VEROSSIMILHANÇA

Modelo	-2LL	Graus de Liberdade	*p-valor*
I	120,04	1	-
II	40,42	10	< 0,0001

FONTE: O autor (2016).

Os resultados demonstrados na tabela 15 indicam que os parâmetros $\beta_1, \ldots, \beta_9$, relativos às variáveis preditoras, são significativos para o modelo. O valor da estatística-teste possui distribuição *Qui-Quadrado* (HOSMER; LEMESHOW, 2000), cujo *p-valor* é inferior ao nível de significância de 5%.

De forma complementar, realizou-se o teste de Wald, conforme TABELA 20.

TABELA 20 - TESTE DE WALD

Variável	*Wald*	*p-valor*
Tamanho	6,9676	0,0083
Margem Operacional	20,2031	< 0,0001
Liquidez Seca	7,1491	0,0075
Participação de Capitais de Terceiros	9,9597	0,0016
Liquidez Corrente	10,6512	0,0011
Prazo Médio de Pagamento de Compras	4,1760	0,041
Giro do Ativo Imobilizado	16,0996	< 0,0001
Giro do Ativo Total	5,1685	0,0230
Composição do Endividamento	6,4653	0,0110

FONTE: O autor (2016).

Em relação aos resultados alcançados no teste de Wald, a TABELA 20 corrobora para a qualidade da estimação dos parâmetros do modelo, uma vez que o *p-valor* indica a significância dos parâmetros estimados, ou seja, rejeita-se a hipótese nula de que os parâmetros da regressão são iguais à zero.

4.2.2.3 Avaliação da qualidade do ajustamento

Para análise da qualidade do ajustamento do modelo, são avaliados os coeficientes Pseudo R^2 de McFadden, R^2 de Cox e Snell e R^2 de Nagelkerke, conforme TABELA 21.

TABELA 21 - MEDIDAS DE ASSOCIAÇÃO MÚLTIPLA

Pseudo R^2	
Cox e Snell	0,5996
Nagelkerke	0,8012
McFadden	0,6633

FONTE: O autor (2016).

Os resultados foram satisfatórios para os três coeficientes apresentados na TABELA 21, o que indica alto poder preditivo do modelo, com boa aderência entre valores observados e previstos.

Hair *et al.*(2010) afirmam que utilização do Pseudo R^2 de McFadden é preferível em relação às demais medidas, e valores entre 0,2 e 0,4 para esta métrica são considerados satisfatórios. Assim, o valor de 0,66 indica elevado poder preditivo do modelo.

No que se refere ao Pseudo R^2 de Cox e Snell, o modelo alcançou um índice de 0,60. Considera-se o resultado bastante satisfatório com base nos outros trabalhos acadêmicos consultados na revisão bibliográfica.

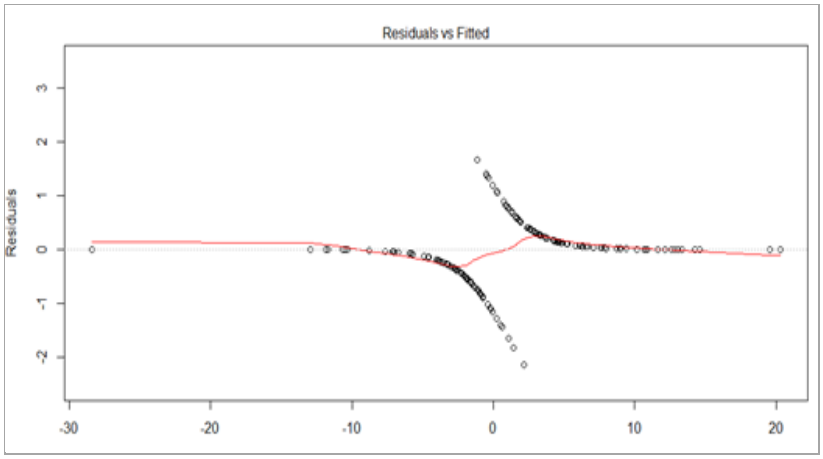
Quanto ao Pseudo R^2 de Nagelkerke, o modelo apresentou um índice de 0,80. Este coeficiente, comparado ao valor máximo de 1,00, corrobora para a alta capacidade classificatória das companhias em estudo.

4.2.2.4 Análise de resíduos

Nesta seção serão analisados alguns aspectos observados na análise dos resíduos, a fim de promover um diagnóstico detalhado do modelo preditivo desenvolvido pela aplicação do método de Regressão Logística.

O GRÁFICO 13 denota a forma com que os resíduos se comportam em relação aos valores preditos. A alocação dos resíduos sobre duas curvas distintas é um padrão esperado, uma vez que a variável resposta do modelo só pode assumir os valores 0 ou 1, enquanto a função logística é não linear e monótona

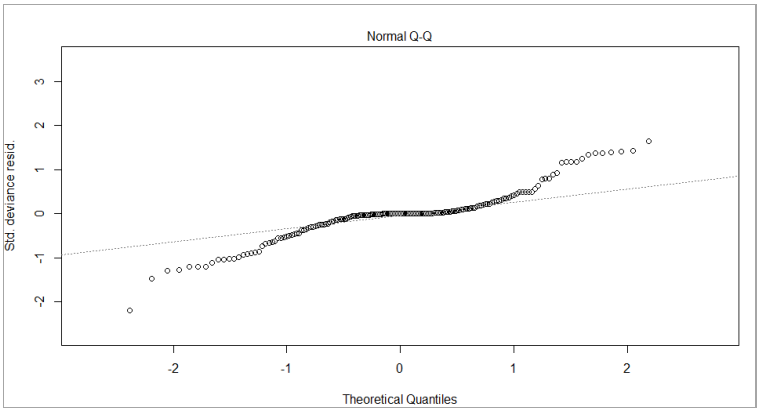
GRÁFICO 13 - RESÍDUOS vs. VALORES PREDITOS.



FONTE: O autor (2016).

Em relação à distribuição de probabilidade dos resíduos do modelo de Regressão Logística, espera-se sua não normalidade (HOSMER; LEMESHOW, 2000), haja vista que a variável resposta do modelo é dicotômica. A ausência de normalidade dos resíduos é comprovada no GRÁFICO 14.

GRÁFICO 14 - QQ-PLOT DOS RESÍDUOS PADRONIZADOS



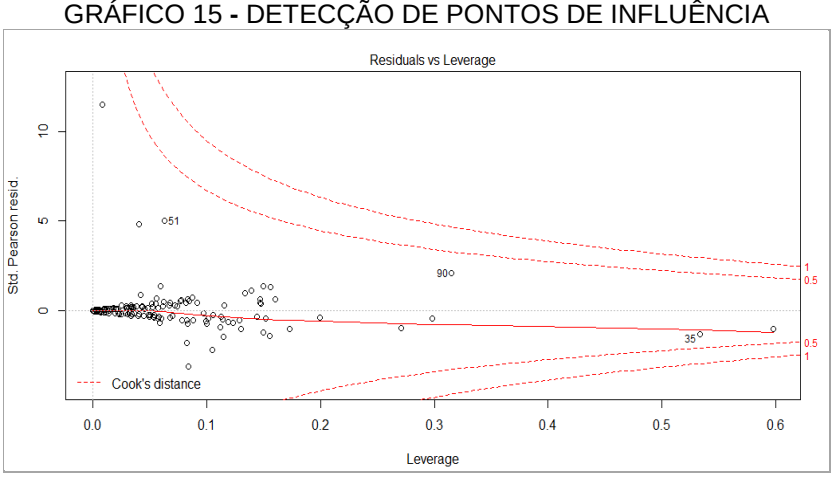
FONTE: O autor (2016).

Uma observação é designada como ponto influente caso possua efeito desproporcional sobre os resultados da regressão, incluindo pontos de alavanca que possuem impacto perceptível nos coeficientes para uma ou mais variáveis independentes (HAIR *et al.*, 2010).

Neste sentido, ressalta-se a importância em se detectar possíveis observações influentes no modelo, as quais podem prejudicar os resultados obtidos pelo modelo de regressão.

O GRÁFICO 15 foi utilizado para análise de possíveis pontos de influência, os quais incluem eventuais pontos de alavanca e os resíduos padronizados.

GRÁFICO 15 - DETECÇÃO DE PONTOS DE INFLUÊNCIA



FONTE: O autor (2016).

O GRÁFICO 15 evidencia a inexistência de pontos além dos limites tracejados no gráfico, o que indica a ausência de pontos de influência significativos.

4.2.2.5 Avaliação de desempenho do modelo

Após validação e diagnóstico do modelo, procedeu-se à análise do desempenho alcançado pela aplicação do método de Regressão Logística ao problema de pesquisa. Analisou-se o poder preditivo do modelo em classificar as empresas da amostra de teste.

A Matriz de Confusão, comparando-se os valores reais e estimados da variável resposta (grupo), permitiu a avaliação do percentual de acertos do modelo, conforme apresentado na TABELA 22.

TABELA 22 - MATRIZ DE CONFUSÃO: REGRESSÃO LOGÍSTICA

GRUPO	*I*	*II*
I	50	6
II	5	57

FONTE: O autor (2016).

O percentual geral de acertos do modelo em classificar as empresas, no ambiente de teste, alcançou a taxa de 90,7%. A classificação realizada pelo modelo obteve os seguintes resultados:

- Classificou-se corretamente 57 companhias com alta rentabilidade;
- Classificou-se corretamente 50 companhias com baixa rentabilidade;
- *Erro Tipo I*: O total de empresas com alta rentabilidade, classificadas incorretamente foi igual a 5;
- *Erro Tipo II*: O total de empresas com baixa rentabilidade, classificadas incorretamente foi igual a 6.

A TABELA 23, por sua vez, revela o percentual de acertos para cada grupo. Nota-se o alto poder preditivo do modelo em classificar tanto empresas de alta rentabilidade como companhias de baixa performance financeira.

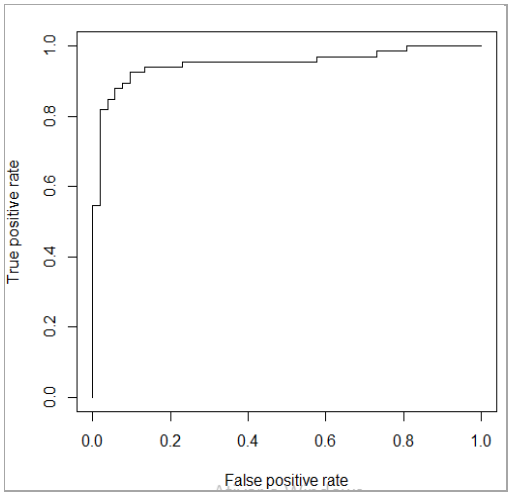
TABELA 23 - PROPORÇÃO DE ACERTOS POR GRUPO

Grupo I	*Grupo II*
89,3%	91,9%

FONTE: O autor (2016).

A área sob a curva ROC apresentada no GRÁFICO 16, alcançou o valor de 0,9432. Assim, de acordo com Hosmer e Lemeshow (2000), considera-se que a aplicação do método de Regressão Logística para reconhecimento de padrões apresentou excelente poder discriminatório.

GRÁFICO 16 - CURVA ROC: REGRESSÃO LOGÍSTICA



FONTE: O autor (2016).

Portanto, nota-se a capacidade do modelo em discriminar as montadoras em dois grupos distintos, os quais sejam:

(i) *Grupo I*: Empresas com baixa rentabilidade; e
(ii) *Grupo II*: Empresas com desempenho superior.

4.3 SELEÇÃO DO MODELO PARA CLASSIFICAÇÃO DE EMPRESAS

Os resultados apresentados na seção anterior refletem de forma satisfatória a previsão do desempenho financeiro de montadoras multinacionais a partir de índices financeiros. Ambos os modelos mostram-se capazes de classificar as

empresas com relação à variável resposta de pertencer ao grupo de alta rentabilidade ou de desempenho inferior.

A TABELA 24 resume os principais resultados alcançados com a aplicação dos métodos estatísticos para reconhecimento de padrões.

TABELA 24 - COMPARAÇÃO ENTRE AS TÉCNICAS

	Percentual de Acertos	Área sob a Curva ROC
Análise Discriminante	89,0%	0,9406
Regressão Logística	**90,7%**	**0,9432**

FONTE: O autor (2016).

Hair *et al.* (2010) apontam algumas razões para utilização da Regressão Logística como alternativa à Análise Discriminante, tais como o menor impacto por conta de desigualdades das matrizes de covariância entre os grupos, maior facilidade de interpretação dos coeficientes e medidas de diagnóstico para análise de resíduos similares à Regressão Múltipla.

Ademais, Hosmer e Lemeshow (2000) afirmam que a Regressão Logística tornou-se um método padrão de análise de regressão para variáveis medidas de forma dicotômica. No caso deste trabalho, a análise de dados envolve a previsão de uma variável com resultado categórico, pois há o interesse em investigar os fatores que determinam a alocação de empresas no grupo de alta rentabilidade, ou seja, quando a variável resposta é igual a 1.

Optou-se pela seleção do método de Regressão Logística para classificação de novas empresas, devido aos melhores resultados alcançados, maior facilidade de interpretação dos resultados e por se tratar de uma abordagem de classificação nos casos em que a variável resposta é dicotômica. Ressalta-se, ainda, que o resultado gerado pela função discriminante linear de Fisher fornece *scores* que possuem menor interpretação intuitiva.

Desta forma, a *logit* estimada para classificação de novas empresas pode ser expressa por (4.1), conforme os coeficientes apresentados na tabela 18:

$$g(\underline{x}) = 8{,}6134 - 0{,}7504x_1 + 0{,}9769x_2 + \cdots + 3{,}7178\,x_8 - 0{,}0523\,x_9 \qquad (4.1)$$

O modelo estimado, por sua vez, pode ser descrito da seguinte forma:

$$\hat{\pi}(\underline{x}) = \frac{e^{8{,}6134-0{,}7504x_1+0{,}9769x_2+\cdots+3{,}7178\,x_8-0{,}0523\,x_9}}{1+e^{8{,}6134-0{,}7504x_1+0{,}9769x_2+\cdots+3{,}7178\,x_8-0{,}0523\,x_9}} \quad (4.2)$$

A TABELA 25 resume as variáveis preditoras e os parâmetros estimados:

TABELA 25 - PARÂMETROS E VARIÁVEIS DA REGRESSÃO LOGÍSTICA

Variável	Descrição	Parâmetro Estimado
-	*Constante*	$\hat{\beta}_0$ = 8,61
x_1	*Tamanho*	$\hat{\beta}_1$ = -0,75
x_2	*Margem Operacional*	$\hat{\beta}_2$ = 0,98
x_3	*Liquidez Seca*	$\hat{\beta}_3$ = 9,54
x_4	*Participação de Capitais de Terceiros*	$\hat{\beta}_4$ = 0,11
x_5	*Liquidez Corrente*	$\hat{\beta}_5$ = -11,71
x_6	*Prazo Médio de Pagamento de Compras*	$\hat{\beta}_6$ = -0,02
x_7	*Giro do Ativo Imobilizado*	$\hat{\beta}_7$ = 0,88
x_8	*Giro do Ativo Total*	$\hat{\beta}_8$ = *3,72*
x_9	*Composição do Endividamento*	$\hat{\beta}_9$ = -0,05

FONTE: O autor (2016).

Para classificação de uma empresa, cada variável preditora é substituída em $\hat{\pi}(x)$, gerando a variável binária que indicará se a empresa pertence ao grupo de baixa rentabilidade (resposta igual à zero) ou alta rentabilidade (resposta igual a 1).

4.3.1 Interpretação das variáveis preditoras

Nesta seção, serão analisadas as variáveis explicativas selecionadas pelo método de Regressão Logística, a fim de fornecer melhor entendimento acerca dos índices que influenciam a rentabilidade das montadoras de veículos.

4.3.1.1 Índices de endividamento

As variáveis “Composição de Endividamento” e “Participação de Capitais de Terceiros” refletem o grau de utilização de recursos de terceiros (financiamentos e empréstimos) para geração de fluxos de caixa das companhias.

No caso das montadoras, a importância destes índices revelou seu caráter de capital intensivo, de tal forma que estes coeficientes possuem importância para medir a saúde financeira geral dessas empresas, e indicam sua capacidade para cumprir as suas obrigações de financiamento.

O parâmetro $\hat{\beta}_9 =$ -0,07, para a Composição do Endividamento, indica que as empresas que possuem maior proporção de dívidas de longo prazo possuem maior chance de serem alocadas no grupo de alta rentabilidade.

Por outro lado, o parâmetro $\hat{\beta}_4$ = 0,09, associado à Participação de Capitais de Terceiros, indica que montadoras mais alavancadas possuem maior chance de auferir melhores resultados financeiros.

4.3.1.2 Giro do ativo imobilizado

Este índice revela a capacidade das montadoras em gerar vendas líquidas a partir de investimentos em imóveis, instalações e equipamentos, como um potencial fator para alta rentabilidade. Portanto, a parcela do faturamento que é investida em bens de capital, surge como variável discriminante para o desempenho financeiro das montadoras.

O parâmetro estimado $\hat{\beta}_7 =$ 1,25 sugere que as montadoras com maior Giro do Ativo Imobilizado possuem maior chance de pertencer ao grupo de empresas de alta rentabilidade.

4.3.1.3 Giro do ativo total

Este coeficiente é utilizado para verificar a capacidade da empresa em gerar vendas, ou seja, indica a maneira pela qual a companhia utiliza seu patrimônio e obtém lucros. O coeficiente $\hat{\beta}_8 =$ 2,37 significa que as montadoras com maior Giro do Ativo Total possuem maior chance de obter maior rentabilidade.

4.3.1.4 Liquidez corrente

No caso das montadoras, este índice é relevante porque as mesmas dispõem de grande diversidade de plantas e alto investimento em instalações.

À medida que há maior investimento em ativos imobilizados, nota-se uma menor liquidez, ou seja, menor capacidade de converter estes ativos em caixa.

O parâmetro $\hat{\beta}_5 =$ -11,05 mostra que as montadoras com menor liquidez corrente têm maior chance de obter alta rentabilidade.

4.3.1.5 Liquidez seca

Maiores níveis de liquidez seca refletem menores níveis de estoques e, portanto, sugerem eficiência da montadora em administrar seus pedidos e estoques, além de indicar o quão rapidamente o estoque existente em seu lote é vendido.

Como o coeficiente desta variável $\hat{\beta}_3$ é igual a 8,35, percebe-se que maiores níveis de liquidez seca podem gerar maior rentabilidade.

4.3.1.6 Margem operacional

Esta variável pode ser interpretada como a parcela do faturamento que é consumida pelas despesas operacionais e pelo custo das mercadorias vendidas. Empresas com baixo custo de mão de obra, por exemplo, possuem margens operacionais mais altas.

O parâmetro $\hat{\beta}_2 = 0{,}92$ indica que as empresas com altas margens operacionais possuem maior chance de figurar no grupo de alta rentabilidade.

4.3.1.7 Prazo médio de pagamento de compras

Este índice revela o tempo que a montadora leva para pagar seus credores, tais como fornecedores. O coeficiente $\hat{\beta}_6 =$ - 0,03 indica maior chance de empresas com menor Prazo Médio de Pagamento de Compras possuírem alta rentabilidade.

4.3.1.8 Tamanho

As montadoras com forte posição no mercado podem operar em diferentes regiões geográficas de todo o mundo e dispõem de estruturas complexas, com

participação em diversos mercados. O parâmetro $\hat{\beta}_1 =$ -1,02 associa o tamanho das empresas à sua rentabilidade.

4.3.2 Aplicação do modelo selecionado

A seguir serão avaliados os resultados do modelo ao classificar cinco empresas selecionadas aleatoriamente do banco de dados. Foram extraídos os índices financeiros referentes aos exercícios contábeis de 2010 a 2014.

Os resultados obtidos pela aplicação da regra de classificação descrita na equação (2.82) podem ser visualizados na TABELA 26.

TABELA 26 **-** APLICAÇÃO DO MODELO SELECIONADO

EMPRESA	T	MO	LS	PCT	LC	PMPC	GAI	GAT	CE	Função Logística	Grupo Predito	Classificação
BMW	25,1	8,4	0,9	78,8	1,1	27,5	5,3	0,5	46,8	**99,8%**	**1**	*Correta*
	25,2	11,7	0,8	78,1	1,0	32,6	6,0	0,4	49,0	**100,0%**	**1**	*Correta*
	25,3	10,8	0,8	77,0	1,0	35,0	6,1	0,4	47,7	**100,0%**	**1**	*Correta*
	25,3	10,4	0,9	74,4	1,0	41,8	2,8	1,2	48,6	**99,9%**	**1**	*Correta*
	25,4	10,8	0,8	76,0	1,0	43,7	2,8	0,5	50,2	**99,6%**	**1**	*Correta*
Dongfeng	23,6	11,9	2,2	57,6	1,3	100,0	6,6	0,4	49,7	**100,0%**	**1**	*Correta*
	23,7	10,9	2,0	55,4	1,3	116,8	6,6	0,4	55,7	**100,0%**	**1**	*Correta*
	23,6	10,1	3,0	53,0	1,4	95,9	5,2	0,5	34,5	**100,0%**	**1**	*Correta*
	22,4	4,1	2,6	45,6	1,2	185,3	2,1	0,1	38,2	**100,0%**	**1**	*Correta*
	23,2	6,0	1,7	49,3	1,0	76,7	7,8	0,2	48,5	**100,0%**	**1**	*Correta*
Honda	25,2	6,4	1,1	61,5	1,3	42,0	2,7	0,7	50,1	**32,8%**	**0**	***Incorreta***
	25,1	2,9	1,0	62,6	1,3	50,4	2,4	0,8	48,5	**1,1%**	**0**	*Correta*
	25,3	5,5	1,0	63,1	1,3	47,2	2,6	0,8	47,6	**16,7%**	**0**	*Correta*
	25,5	6,3	0,9	62,1	1,2	42,2	2,6	0,9	48,6	**34,0%**	**0**	*Correta*
	25,6	5,0	0,9	61,4	1,2	39,4	2,3	0,9	46,8	**13,0%**	**0**	*Correta*
Mazda	23,9	1,0	1,0	75,8	1,3	50,9	2,9	1,8	47,8	**56,7%**	**1**	***Incorreta***
	23,7	-1,9	1,2	75,5	1,6	54,7	2,6	1,7	43,0	**1,4%**	**0**	*Correta*
	23,8	2,5	1,0	74,9	1,4	60,1	2,8	1,6	51,2	**44,0%**	**0**	*Correta*
	24,0	6,8	1,0	70,6	1,4	61,5	3,3	1,3	51,2	**89,8%**	**1**	*Correta*
	24,1	6,7	1,0	64,8	1,5	64,2	3,4	1,1	56,4	**49,4%**	**0**	*Correta*
Renault	24,7	1,6	0,9	68,3	1,0	71,2	3,3	0,5	77,6	**0,9%**	**0**	***Incorreta***
	24,7	2,9	0,9	67,0	1,0	65,9	3,7	0,5	79,7	**3,6%**	**0**	*Correta*
	24,7	0,3	0,9	67,8	1,0	68,3	3,6	0,5	79,8	**0,3%**	**0**	*Correta*
	24,7	-0,1	1,0	69,6	1,1	69,1	3,6	0,5	78,4	**0,3%**	**0**	*Correta*
	24,7	2,7	1,0	70,0	1,1	72,7	3,8	0,4	78,6	**4,7%**	**0**	*Correta*

FONTE: O autor (2016).

Os índices financeiros estão apresentados sob a seguinte forma:

- T = Tamanho;
- MO = Margem Operacional;
- LS = Liquidez Seca;
- PCT = Participação de Capitais de Terceiros;
- LC = Liquidez Corrente;

- PMPC = Prazo Médio de Pagamento de Compras;
- GAI = Giro do Ativo Imobilizado;
- GAT = Giro do Ativo Total;
- CE = Composição do Endividamento.

A função logística $\hat{\pi}(\underline{x})$ indica a estimativa da probabilidade de que a respectiva empresa, com o vetor de variáveis explicativas $\underline{x}$, possua alta rentabilidade frente a seus concorrentes e, portanto, seja alocada no grupo 2.

Nota-se o alto poder preditivo do modelo ao classificar as montadoras, alcançando o percentual de acertos de 88% nos exemplos apresentados.

4.3.3 Desenvolvimento de ferramenta para simulação de cenários

Com a finalidade de elucidar o funcionamento do modelo selecionado para análise financeira das montadoras de veículos, optou-se pelo desenvolvimento de um simulador no aplicativo *Microsoft Excel*, com utilização da linguagem de programação *Visual Basic for Applications* (VBA).

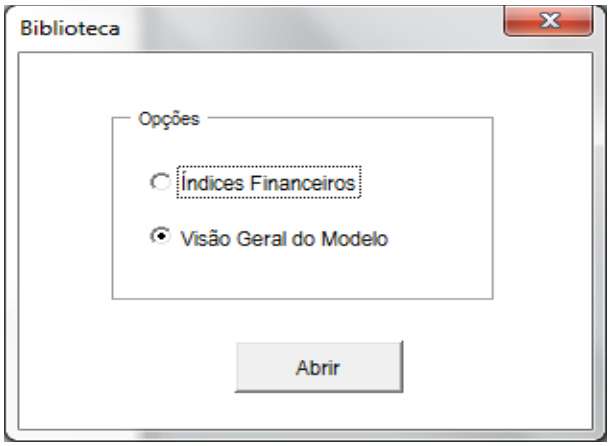
Este simulador visa proporcionar a construção de diferentes cenários para a rentabilidade das montadoras, em função de seus índices financeiros. A FIGURA 16 exibe a tela inicial da ferramenta.

FIGURA 16 - APRESENTAÇÃO DO SIMULADOR DE RENTABILIDADE

FONTE: O autor (2016).

Conforme apresentado na FIGURA 17, o módulo “Biblioteca” consiste na exposição do arcabouço teórico referente aos índices financeiros utilizados como variáveis preditoras do modelo, bem como o funcionamento da metodologia empregada para classificação das empresas.

FIGURA 17 - BIBLIOTECA DO SIMULADOR DE RENTABILIDADE



FONTE: O autor (2016).

No que tange às variáveis preditoras do nível de rentabilidade das empresas, a opção “Índices financeiros” traz a formulação das variáveis utilizadas pelo modelo à luz da literatura acadêmica, conforme disposto na FIGURA 18.

FIGURA 18 - ÍNDICES DO SIMULADOR DE RENTABILIDADE

FONTE: O autor (2016).

Ademais, o módulo de biblioteca oferece a opção “Visão Geral do Modelo”, proporcionando ao usuário melhor compreensão acerca dos métodos empregados e do modelo estimado.

Na tela apresentada na FIGURA 19, encontra-se a descrição do modelo preditivo utilizado pelo simulador, seus objetivos e suas bases técnicas, bem como os parâmetros estimados e os grupos de rentabilidade nos quais as empresas poderão ser alocadas.

FIGURA 19 - MODELO UTILIZADO NO SIMULADOR DE RENTABILIDADE

FONTE: O autor (2016).

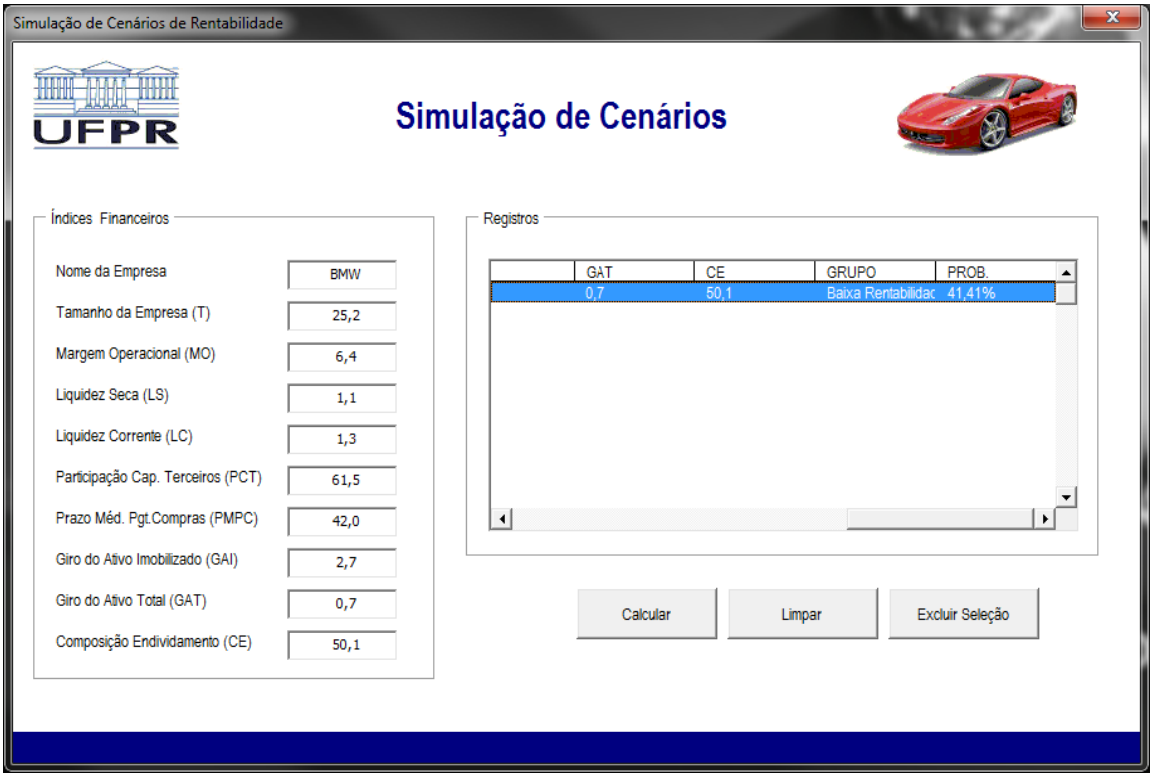
De forma a proporcionar um canal de suporte ao usuário, optou-se pelo desenvolvimento do módulo “Contatos”, apresentado na FIGURA 20. De tal modo, em caso de dúvidas técnicas sobre o funcionamento do modelo ou aspectos gerais da ferramenta, o usuário poderá consultar os pesquisadores responsáveis pelo desenvolvimento do simulador.

FIGURA 20 - CONTATOS SOBRE O SIMULADOR DE RENTABILIDADE

FONTE: O autor (2016).

O módulo principal da ferramenta pode ser acessado pelo botão “Clique Aqui para Iniciar” da tela de apresentação, conforme FIGURA 21. Neste módulo, o usuário deve inserir os índices financeiros da empresa, obtidos através de projeções financeiras, relatórios, orçamentos ou outras fontes de informação.

FIGURA 21 - TELA PRINCIPAL DO SIMULADOR DE RENTABILIDADE



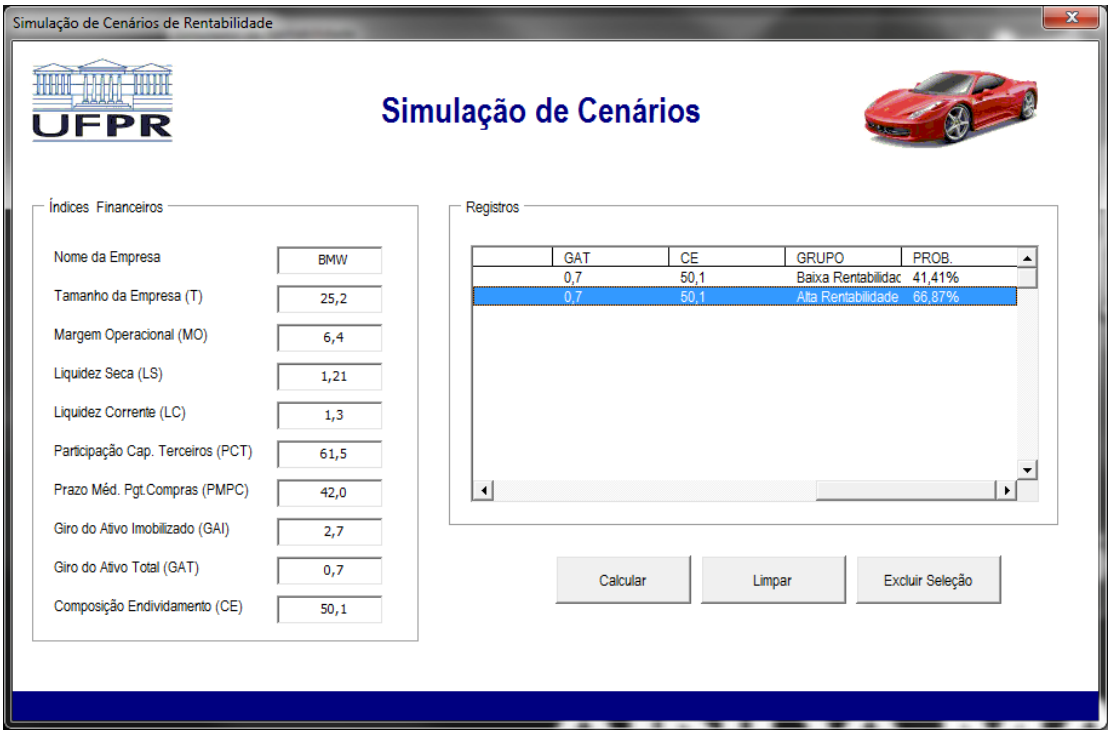
FONTE: O autor (2016).

Com base nos índices financeiros, o modelo informa se a empresa é alocada no grupo de alta ou de baixa rentabilidade. A variável “PROB” indica a probabilidade

de uma companhia ser classificada no grupo de alta rentabilidade e, caso esta probabilidade seja menor que 50%, a empresa é classificada no grupo de baixa rentabilidade. No exemplo da FIGURA 21, o usuário realizou uma simulação para determinada montadora, a qual possui uma probabilidade de 41,4% de ser alocada no grupo de alta rentabilidade.

A FIGURA 22, por sua vez, traz um exemplo de uma simulação de cenários, onde o usuário deseja analisar a rentabilidade da mesma empresa no caso de um aumento de 10% em seu índice de Liquidez Seca, por conta de uma nova política de redução de estoques. Neste caso, com um aumento de 10% sobre o índice de Liquidez Seca, mantendo-se os demais índices constantes, observou-se que a probabilidade de que a empresa possua alta rentabilidade aumentou para 66,87%.

FIGURA 22 - APLICAÇÃO DO SIMULADOR DE RENTABILIDADE



FONTE: O autor (2016).

Em linhas gerais, o simulador tem a finalidade de servir como um instrumento relevante para a análise financeira das montadoras de veículos, pois se torna possível analisar as variáveis que possuem influência sobre a rentabilidade destas companhias, além de permitir a definição de metas para seus índices financeiros em função dos resultados desejados para os negócios.

5 CONSIDERAÇÕES FINAIS

O panorama da indústria automotiva enfrenta mudanças significativas ao longo das últimas décadas, com a rápida expansão em mercados emergentes, estimulada por incentivos governamentais e estratégias de liderança em custos. Desta forma, os fabricantes de automóveis podem buscar subsídios para desenvolvimento de estratégias de produção que permitam reduzir seus custos e aumentar sua competitividade.

Neste sentido, a análise financeira de empresas do setor automotivo configura-se como uma área de particular importância, haja vista a alta competitividade do setor e sua influência sobre a economia global. Restrições quanto à obtenção de dados financeiros destas companhias, porém, podem surgir como barreiras para a elaboração de trabalhos acadêmicos.

Os dados financeiros coletados neste trabalho permitiram a aplicação de métodos multivariados para análise financeira de montadoras multinacionais, apontando-se as variáveis relevantes para o nível de rentabilidade destas empresas. Os métodos de Análise Discriminante de Fisher e Regressão Logística alcançaram alto poder preditivo e resultados satisfatórios quanto à alocação das empresas nos grupos definidos através da Análise de Agrupamentos.

As variáveis discriminantes selecionadas por ambos os métodos refletem aspectos técnicos inerentes às empresas do setor automotivo, tais como utilização de recursos de terceiros para investimentos de grande porte e investimentos em bens de capital.

O método de Regressão Logística foi selecionado para classificação de empresas devido a vantagens decorrentes de suas propriedades e por apresentar resultados superiores. Aproximadamente 91% das empresas foram classificadas corretamente pelo método no ambiente de testes, alcançando a área sob a curva ROC de aproximadamente 0,94. À luz da literatura acadêmica, consideram-se os resultados bastante satisfatórios.

Todos os resultados são baseados na classificação gerada pela técnica de *clustering*. Além disto, as avaliações utilizaram dados do período entre os anos de 2006 e 2015, de modo que a atualização da calibragem das variáveis preditoras e reconhecimento de novos padrões demanda a atualização do banco de dados.

Este trabalho proporcionou a compreensão prática de técnicas multivariadas, reforçando o conhecimento teórico e ratificando a aplicabilidade dos métodos à análise da situação financeira de montadoras de veículos. Assim, as companhias em posse de orçamentos anuais ou projeções, possuem condições de prever seu nível de rentabilidade sob diferentes cenários. Neste sentido, a experiência empírica do analista e a robustez do modelo podem ajudar na tarefa de tomada de decisão.

Sob a ótica social, este trabalho concede subsídios para decisões estratégicas de montadoras de veículos e, portanto, contribui para maior solidez destas companhias e estimula a manutenção de seu nível de empregabilidade. De igual modo, o conhecimento das variáveis que determinam o diferencial competitivo para estas companhias, reforça sua imagem juntos aos credores, investidores e demais *stakeholders.*

Ademais, vislumbra-se a possibilidade de futuros trabalhos acadêmicos de natureza similar. Sugere-se o desenvolvimento de pesquisas com a finalidade de comparar os métodos multivariados empregados neste trabalho com outras metodologias empregadas para reconhecimento de padrões, tais como técnicas de Inteligência Artificial e metaheurísticas.

Os excelentes resultados obtidos neste trabalho indicam a possibilidade da aplicação de modelos multivariados à análise financeira de montadoras de veículos. Neste contexto, a utilização de modelos quantitativos pode oferecer contribuições relevantes para mitigação dos riscos gerenciáveis incorridos nos negócios destas companhias e exercer impactos positivos para sua sustentabilidade.

REFERÊNCIAS

AHN, B.S.; C.H.O.; S.S.; KIM, C.Y. The integrated methodology of rough set theory and artificial neural network for business failure prediction. **Expert Systems With Applications**, v.18, 2000.

ALDENDERFER, M.S.; BLASHFIELD, R.K. **Cluster Analysis.** Sage University Paper Series: Quantitative Applications in the Social Science. New York, 1984.

ALTMAN, E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, **The Journal of Finance**, 1968.

ALTMAN, E. **Corporate financial distress**: A complete Guide do predicting, avoiding and dealing with bankruptcy, 2nd edition. John Wiley and Sons. New York, 1993.

ANFAVEA - Associação Nacional dos Fabricantes de Veículos Automotores. Disponível em: <http://www.anfavea.com.br>. Acesso em: 10/10/2015.

ANTE, A.; ANA, K. Discriminant Analysis of Bank Profitability Levels. **Croatian Operational Research Review**, Vol. 4. Croatia, 2013.

ASSAF NETO, A. **Finanças Corporativas e Valor**, 3ª edição. Atlas. São Paulo, 2007.

ATIYA, AF. Bankruptcy prediction for credit risk using neural networks: A survey and new results. **IEEE Transactions on Neural Networks**, v.12, 2001.

BOYACIOGLU, M.A.; KARA, Y.; BAYKAN, O.K. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods. **Expert Systems With Applications**, v.36, 2009.

BREALEY, R.A.; MYERS, S.C.; MARCUS, A.J. **Fundamentals of Corporate Finance**. 3rd edition. McGraw Hill Companies, Inc. New York, 2001.

BRITO, G.A.S.; ASSAF NETO, A. Modelo de classificação de risco de credito de empresas. **Revista Contabilidade & Finanças**, v. 46, 2008.

BUSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à Análise de Agrupamentos**. Associação Brasileira de Estatística. São Paulo, 1999.

DONOVAN, C.; NUÑES, L. **Figuring what's fair**: The cost of equity capital for renewable energy in emerging markets. Elsevier. Madrid, 2010.

FÁVERO, L.P.; BELFIORE, P.; SILVA, F.L.; CHAN, B.L. **Análise de dados - Modelagem multivariada para tomada de decisões**. Elsevier. Rio de Janeiro, 2009.

FENABRAVE - Federação Nacional da Distribuição de Veículos Automotores. **Anuário do Desempenho distribuição automotiva no Brasil**. São Paulo, 2014.

FENABRAVE - Federação Nacional da Distribuição de Veículos Automotores. Disponível em: <http://www.fenabrave.org.br>. Acesso em: 08/11/2015.

FONTALVO, H.T.; J.; GRANADILLO, E.; VERGARA, J.C. Application of discriminant analysis to evaluate the improvement of financial indicators in the food sector companies Barranquilla-Colombia. **Revista chilena de Ingeniería**, v.20, 2012.

GIL, A.C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GITMAN, L.J.; MADURA, J. **Administração financeira**: uma abordagem gerencial. Addison Wesley. São Paulo, 2003.

HAIR, J.F.; ANDERSON, R.E.; TATHANM, R.L.; BLACK, W.C. **Multivariate Data Analysis**, 17th. Ed., Pearson Prentice Hall, New Jersey, 2010.

HOSMER, D.W.; LEMESHOW S.; **Applied Logistic Regression**. 2nd Ed., John Wiley and Sons Inc. New York, 2000.

JAIN, A. K.; DUIN, R.P.W.; MAO, J. Statistical pattern recognition: A review**. IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol 22. 2000.

JARILLO J. C. **Strategic Logic**. Palgrave McMillan. New York, 2003.

JOHNSON, R.A.; WICHERN, D.W. **Applied Multivariate Statistical Analysis**. Prantice Hall. New Jersey, 2007.

KANITZ, S. C. **Controladoria**: Teoria e estudos de casos. Editora Pioneira. São Paulo, 1976.

KENDALL, M. G. **Multivariate Analysis**. 2. ed. High Wycombe: Charles Griffin. London ,1980.

KIRK, J.; MILLER, M.L.; **Reliability and validity in qualitative research**. Sage. Beverly Hills, 1986.

MARION, J.C.M. **Análise das Demonstrações Contábeis**: Contabilidade Empresarial, 3ª edição. Editora Atlas. São Paulo, 2007.

MARRIOT, F.H.C. **The interpretation of multivariate observations**. London, Academic Press, 1974.

MARTINS, G.A. **Estatística Geral e Aplicada**. Atlas. São Paulo, 2001.

MATARAZZO, D. C. **Análise Financeira de Balanços**: Abordagem Básica e Gerencial. 6ª Edição. Editora Atlas. São Paulo, 2010.

MDIC – Ministério do Desenvolvimento, Indústria e Comércio Exterior. Disponível em: <http://www.mdic.gov.br>. Acesso em: 01/11/2015.

MIN, S.H.; LEE, J.; Han, I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. **Expert Systems With Applications**, v.31, 2006.

OH, M.K. Financial Distress Prediction Models for Wind Energy SMEs. **International Journal of Contents**, v.10, 2014.

PADOVEZE, C.L. **Contabilidade gerencial**: um enfoque em sistema de informação contábil. 5ª Edição. Editora Atlas. São Paulo, 2008.

PEREIRA, J.C.R. **Análise de dados qualitativos**. Editora da Universidade de São Paulo/FAPESP. São Paulo, 1999.

PORTER, M.E. **Estratégia competitiva**: técnicas para análise da indústria e da concorrência. 7ª edição. Rio de Janeiro, 1986.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Disponível em: < http://www.R-project.org/>. Acesso em 10/03/2015.

REIS, E. **Estatística Multivariada Aplicada**, 2ª edição. Editora Sílabo. Lisboa, 2001.

RIBEIRO, O. M. **Estrutura e análise de balanço fácil**. 8ª Edição. Editora Saraiva. São Paulo, 2009.

ROSS, S.A.; WESTERFIELD, R.W.; JAFFE, J.F. **Administração Financeira**: Corporate Finance, 2ª edição. Editora Atlas. São Paulo, 2002.

SCARPEL, R.A. **Modelos de previsão de insolvência**: uma abordagem discriminante paramétrica e não paramétrica. ITA/IMB. São José dos Campos, 2005.

SHIN, K.S.; LEE, T.S.; KIM, H.J. An application of support vector machines in bankruptcy prediction model. **Expert Systems With Applications**, v.28, 2005.

SILVA, J.P. **Análise financeira das empresas**. 9ª Edição. Editora Atlas. São Paulo, 2008.

SLACK, N.; CHAMBERS, S.; JOHNSTON, R. **Administração da Produção**. 2ª Edição. Editora Atlas. São Paulo, 2002.

STERN, L.W.; EL-ANSARY, A.; COUGHLAN, A.T. **Marketing Channels**. 5th edition. Prentice Hall. New Jersey, 1996.

STEVENSON, W.J. **Administração das Operações de Produção**. 6ª edição. Editora LTC. Rio de Janeiro, 2001.

TSERNG, H.P.; CHEN, P.C.; HUANG, W.H.; LEI, M.C.; TRAN, Q.H.; HAN, I. Prediction Of Default Probability For Construction Firms Using The Logit Model. **Journal of Civil Engineering and Management**, v.20, 2014.

VAN HORNE, J.C; WACHOWICZ J.M. **Fundamentals of Financial Management**. 13th Edition. Pearson Education Limited. London, 2008.

WERNKE, R. **Gestão Financeira**: Ênfase em Aplicações e Casos Nacionais. Editora Saraiva. Rio de Janeiro, 2008.

WHITE, G.I.; SONDHI A.C.; FRIED, D. **The Analysis and Use of Financial Statements**, 3rd Edition. John Wiley & Sons, Inc. New York, 2002.

APÊNDICE 1

Códigos utilizados no software R - Análise de Agrupamentos

```
#### CARREGANDO BASE DE DADOS
RATIOS<-read.csv2(file='Base_Clustering.csv')
RATIOS

#### CRIANDO UMA MATRIZ DE DISTÂNCIAS
d <- dist(RATIOS[2:4], method="euclidian")
d

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - LIGAÇÃO SIMPLES
   *********************************************\n")
cl.s <- hclust(d,method = "simple")
plot(cl.s, main="Dendrograma - Ligacao Simples", hang=-1)

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - LIGAÇÃO COMPLETA
   *********************************************\n")
cl.c <- hclust(d, method = "complete")
plot(cl.c, main="Dendrograma - Ligacao Completa", hang=-1)

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - LIGAÇÃO MÉDIA
   *********************************************\n")
cl.m <- hclust(d, method = "average")
plot(cl.m, main="Dendrograma - Ligacao Média",hang=-1)

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - MÉTODO DO CENTRÓIDE
   *********************************************\n")
cl.ce <- hclust(d, method = "centroid")
plot(cl.ce, main="Dendrograma - Centroid Linkage",hang=-1)

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - LIGAÇÃO MEDIANA
   *********************************************\n")
cl.me <- hclust(d, method = "median")
plot(cl.me, main="Dendrograma - Median Linkage",hang=-1)

cat("    *********************************************
     AGRUPAMENTO HIERARQUICO - MÉTODO DE WARD
```

```
   ***********************************************\n")
cl.w <- hclust(d, method = "ward.D2")
plot(cl.w, main="Dendrograma - Ward's Linkage",hang=-1)

######## FORMANDO DOIS CLUSTERS
cat("    ----------------------------------------------------------------
    Resumo da Quantidade de Observações nos Cluster Selecionados
   ---------------------------------------------------------------- \n")
groups.2 = cutree(cl.w,2)
table(groups.2)

cat("    ----------------------------------------------------------------
    Agrupamento das Empresas Selecionadas conforme lucratividade
   ---------------------------------------------------------------- \n")
sapply(unique(groups.2),function(g)RATIOS$ROE[groups.2 == g])

######## GERANDO NOVA BASE COM OS CLUSTERS E CALCULANDO A MANOVA

cat("    ----------------------------------
            Base de Dados Agrupada
   ---------------------------------- \n")
CLUSTER <- groups.2
MATRIZ_CLUSTER <- cbind(RATIOS,CLUSTER)
MATRIZ_CLUSTER

grupos <- MATRIZ_CLUSTER$CLUSTER

Y = cbind(MATRIZ_CLUSTER$ROA,MATRIZ_CLUSTER$ROE,MATRIZ_CLUSTER$ROIC )

### AVALIANDO A QUALIDADE DO CLUSTERING

MANO <- manova(Y~grupos)

RES_MAN <- summary(MANO, test="Wilks")
RES_MAN2 <- summary(MANO, test="Pillai")
RES_MAN3 <- summary(MANO, test="Hotelling-Lawley")
RES_MAN4 <- summary(MANO, test="Roy")

cat("    ******************
     LAMBDA DE WILKS
   ******************\n")
RES_MAN

cat("    ******************
     TRAÇO DE PILLAI
```

```
   ********************\n")
RES_MAN2

cat("    ******************
     HOTTELING-LAWLEY
   ********************\n")
RES_MAN3

cat("    ******************
     RAIZ MÁXIMA DE ROY
   ********************\n")
RES_MAN4
```

APÊNDICE 2

Códigos utilizados no software R - Análise Discriminante e Regressão Logística

```
############ ANALISE DISCRIMINANTE#######################

#**************** BASE DE DADOS DE TREINAMENTO ******************#

RECPAD<-read.csv2(file='AMOSTRA_TREINAMENTO.csv')
RECPAD
RECPAD2 <- RECPAD[,5:20]
RECPAD2

require(MASS)
ANALISE <- lda(formula = GRUPO ~ .,
                        data = RECPAD2)

#### SELECIONANDO VARIAVEIS COM BASE NO LAMBDA WILKS
require(klaR)
gw_obj = greedy.wilks(GRUPO ~ ., data=RECPAD2)
gw_obj

ANALISE_STEP <- lda(formula = GRUPO ~ MARGEM_OP + GIRO_ATIVOS + ENDIV_LP + IMOB_PL +
LIQ_CORRENTE +
   LIQ_IMED + GIRO_ESTOQUE + TAMANHO + GIROT_AT_FIXOS + PMPC ,
                        data = RECPAD2)

cat("    ******************************
     VALIDAÇÃO HIPÓTESES DO MODELO
   ******************************
   SERAO TESTADAS CADA VARIAVEL DO MODELO:

      1)VARIAVEIS IDENDEPENDENTES ~ NORMAL : ** FDL FISHER DISPENSA
      2)GRUPOS COM MATRIZ DE COVARIANCIAS HOMOGENEAS
      3)LINEARIDADE DAS VARIAVEIS INDEPENDENTES
      4)AUSENCIA DE COLINEARIDADE

 #### TESTES DE HOMOGENEIDADE DE MATRIZES DE COVARIANCIA ####

bartlett.test(RECPAD2$MARGEM_OP,RECPAD2$GRUPO)
bartlett.test(RECPAD2$GIRO_ATIVOS,RECPAD2$GRUPO)
bartlett.test(RECPAD2$ENDIV_LP,RECPAD2$GRUPO)
bartlett.test(RECPAD2$IMOB_PL,RECPAD2$GRUPO)
bartlett.test(RECPAD2$LIQ_CORRENTE,RECPAD2$GRUPO)
```

```
bartlett.test(RECPAD2$LIQ_IMED,RECPAD2$GRUPO)
bartlett.test(RECPAD2$GIRO_ESTOQUE,RECPAD2$GRUPO)
bartlett.test(RECPAD2$TAMANHO,RECPAD2$GRUPO)
bartlett.test(RECPAD2$GIROT_AT_FIXOS,RECPAD2$GRUPO)
bartlett.test(RECPAD2$PMPC,RECPAD2$GRUPO)

#######   GRÁFICOS BOX PLOT   ##########

par(mfrow=c(3,4))
boxplot(RECPAD2$MARGEM_OP ~ RECPAD2$GRUPO, main="MARGEM OPERACIONAL",
                              xlab="GRUPO" )
boxplot(RECPAD2$GIRO_ATIVOS ~ RECPAD2$GRUPO, main="GIRO DO ATIVO TOTAL",
                              xlab="GRUPO" )
boxplot(RECPAD2$ENDIV_LP ~ RECPAD2$GRUPO, main="ENDIVIDAMENTO DE LONGO PRAZO",
                              xlab="GRUPO" )
boxplot(RECPAD2$IMOB_PL ~ RECPAD2$GRUPO, main="IMOBILIZAÇÃO SOBRE PL",
                              xlab="GRUPO" )
boxplot(RECPAD2$LIQ_CORRENTE ~ RECPAD2$GRUPO, main="LIQUIDEZ CORRENTE",
                              xlab="GRUPO" )
boxplot(RECPAD2$LIQ_IMED ~ RECPAD2$GRUPO, main="LIQUIDEZ IMEDIATA",
                              xlab="GRUPO")
boxplot(RECPAD2$GIRO_ESTOQUE ~ RECPAD2$GRUPO, main="GIRO DE ESTOQUE",
                              xlab="GRUPO" )
boxplot(RECPAD2$TAMANHO ~ RECPAD2$GRUPO, main="TAMANHO",
                              xlab="GRUPO" )
boxplot(RECPAD2$GIROT_AT_FIXOS ~ RECPAD2$GRUPO, main="GIRO DO ATIVO FIXO",
                              xlab="GRUPO")
boxplot(RECPAD2$PMPC ~ RECPAD2$GRUPO, main="PRAZO MÉDIO DE PAGAMENTO DE COMPRAS",
                             xlab="GRUPO" )

######## GRÁFICOS DISPERSÃO ##########

yyy <- cbind(RECPAD2$ENDIV_LP, RECPAD2$LIQ_IMED, RECPAD2$TAMANHO, RECPAD2$PMPC)

cols <- character(nrow(RECPAD2))
cols[] <- "black"
cols[RECPAD2$GRUPO == 0] <- "darkred"
cols[RECPAD2$GRUPO == 1] <- "darkblue"
pairs(yyy, main="Gráfico de Dispersão", pch=22,col=cols[])

cat("    ***************************
     PROBABILIDADES A PRIORI
   ***************************\n")
ANALISE_STEP$prior
```

```
cat("    ***********************************
     CONTAGEM DE ELEMENTOS NOS GRUPOS
   ***********************************\n")
ANALISE_STEP$counts

cat("    *******************************
     MÉDIA DAS VARIÁVEIS POR GRUPO
   *******************************\n")
ANALISE_STEP$means

cat("    ****************************************
     FUNÇÃO DISCRIMINANTE LINEAR DE FISHER
   ****************************************\n")
ANALISE_STEP$scaling
 ANALISE_STEP$svd

PROP = ANALISE_STEP$svd^2/sum(ANALISE_STEP$svd^2)
PROP

##### CLASSIFICAÇÃO DAS EMPRESAS
pred <- predict(ANALISE_STEP)
pred$class

cat("    **********************
     MATRIZ DE CONFUSÃO
   **********************\n")

tabela <- table(RECPAD2$GRUPO, pred$class)
tabela

cat("    **********************
     PROPORÇÃO DE ACERTOS
   **********************\n")
diag(prop.table(tabela,1))
sum(diag(prop.table(tabela)))

cat("    ********************************************
     ANÁLISE DA CURVA ROC - AMOSTRA TREINAMENTO
   ********************************************\n")

p_lda <- predict(ANALISE_STEP)$post[,2]
pr_lda <- prediction(p_lda, RECPAD2$GRUPO)
prf_lda <- performance(pr_lda, "tpr", "fpr")
plot(prf_lda)
```

```
auc <- performance(pr_lda, measure = "auc")
auc <- auc@y.values[[1]]
auc

#*************** BASE DE DADOS DE TESTE ************************************#

RECPAD_TEST1 <-read.csv2(file='AMOSTRA_TESTE.csv')
RECPAD_TEST1

RECPAD_TEST <- RECPAD_TEST1[,5:20]
RECPAD_TEST

pred2 <- predict(ANALISE_STEP, newdata=RECPAD_TEST)
pred2$class

cat("    **********************
     MATRIZ DE CONFUSÃO
   **********************\n")

tabela7 <- table(RECPAD_TEST$GRUPO, pred2$class)
tabela7

cat("    **********************
     PROPORÇÃO DE ACERTOS
   **********************\n")

sum(diag(tabela7)) / sum(tabela7)

cat("    ***************************************
     ANÁLISE DA CURVA ROC - AMOSTRA TESTES
   ***************************************\n")

p_lda <- predict(ANALISE_STEP2, newdata=RECPAD_TEST)$post[,2]
pr_lda <- prediction(p_lda, RECPAD_TEST$GRUPO)
prf_lda <- performance(pr_lda, "tpr", "fpr")
plot(prf_lda)

auc <- performance(pr_lda, measure = "auc")
auc <- auc@y.values[[1]]
auc

cat("    **********
     MANOVA
   **********\n")
```

```
GRUP <- RECPAD_TEST$GRUPO

GR2 <-
cbind(RECPAD_TEST$MARGEM_OP,RECPAD_TEST$GIRO_ATIVOS,RECPAD_TEST$ENDIV_LP,RECPAD_
TEST$IMOB_PL,
RECPAD_TEST$LIQ_CORRENTE,RECPAD_TEST$LIQ_IMED,
RECPAD_TEST$GIRO_ESTOQUE,RECPAD_TEST$TAMANHO,
RECPAD_TEST$GIROT_AT_FIXOS,RECPAD_TEST$PMPC)

AV <- manova(GR2 ~ GRUP)

RES_MAN11 <- summary(AV, test="Wilks")
RES_MAN12 <- summary(AV, test="Pillai")
RES_MAN13 <- summary(AV, test="Hotelling-Lawley")
RES_MAN14 <- summary(AV, test="Roy")

cat("    ******************
     LAMBDA DE WILKS
   ******************\n")
RES_MAN11

cat("    ******************
     TRAÇO DE PILLAI
   ******************\n")
RES_MAN12

cat("    ******************
     HOTTELING-LAWLEY
   ******************\n")
RES_MAN13

cat("    ******************
     RAIZ MÁXIMA DE ROY
   ******************\n")
RES_MAN14



############ REGRESSÃO LOGÍSTICA########################


#************************ BASE DE DADOS DE TREINAMENTO ****************#

RECPAD<-read.csv2(file='AMOSTRA_TREINAMENTO.csv')
RECPAD
```

```
RECPAD2 <- RECPAD[,5:20]
RECPAD2

mylogit <- glm(GRUPO ~ .,family = "binomial"(link='logit'), data=RECPAD2)
summary(mylogit)

cat("    ********************************************
     MÉTODO STEPWISE PARA SELEÇÃO DAS VARIÁVEIS
   ********************************************\n")
step(mylogit)

##### DEFININDO MODELO COM AS VARIÁVEIS SELECIONADOS NO STEPWISE

mylogit2 <-  glm(formula = GRUPO ~ TAMANHO + MARGEM_OP + LIQ_SECA + PART_CAP_TERC +
   LIQ_CORRENTE + PMPC + GIROT_AT_FIXOS + GIRO_ATIVOS + COMPOSIC_END,
   family = binomial(link = "logit"), data = RECPAD2)

cat("    **************************
     PROPORÇÃO DE ACERTOS
   **************************\n")

tabelaxx <- table(RECPAD2$GRUPO, p>0.50)
sum(diag(tabelaxx)) / sum(tabelaxx)

cat("    ********************************
     ANÁLISE DA CURVA ROC - TREINAMENTO
   ********************************\n")
require(ROCR)

p <- predict(mylogit2, type = "response")
pr <- prediction(p, RECPAD2$GRUPO)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

#************************************ BANCO DE DADOS DE TESTE ***************************

RECPAD_TEST1 <-read.csv2(file='AMOSTRA_TESTE.csv')
RECPAD_TEST1
RECPAD_TEST <- RECPAD_TEST1[,5:20]
RECPAD_TEST
```

```
cat("    ******************************************
     ANÁLISE DA CURVA ROC - AMOSTRA DE TESTE
   ******************************************\n")

require(ROCR)

p <- predict(mylogit2,newdata=RECPAD_TEST,type = "response")
pr <- prediction(p, RECPAD_TEST$GRUPO)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)

auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

cat("    **************************
     PROPORÇÃO DE ACERTOS
   **************************\n")

tabela99 <- table(RECPAD_TEST$GRUPO, p>0.50)
sum(diag(tabela99)) / sum(tabela99)

cat("    ************************
     ANÁLISE DE RESÍDUOS
   ************************\n")
plot(residuals(mylogit2))
par(mfrom=c(1,1))
plot(mylogit2)
layout(1)

cat("    *********************************
     RAZÃO DE VEROSSIMILHANÇA
   *********************************\n")
require (lrtest)
MODELO_SEM_COEF <- glm(RECPAD2$GRUPO ~1, family = "binomial")
RV <- lrtest(MODELO_SEM_COEF,mylogit2)
RV

cat("    ******************
     TESTE DE WALD
   ******************\n")

require(AOD)
WALD <- wald.test(b=coef(mylogit2), vcov(mylogit2),Term = 1:9 )
WALD
```

```
cat("    ************************************
     MEDIDAS DE ASSOCIAÇÃO MULTIPLA
   ************************************\n")

require(pscl) #Nagelkerke = CU; Cox & Snell = ML
pR2(mylogit2)

cat(" *   R2CU = PSEUDO R² DE NAGELKERKE
     **  R2ML = PSEUDO R²DE COX & SNELL \n")
```

APÊNDICE 3

Dados Coletados e Classificação das Empresas - Amostra de Testes

EMPRESA	T	MO	LS	PCT	LC	PMPC	GAI	GAT	CE	Grupo *Clustering*	Grupo Predito	Classificação
AVTOVAZ	22,3	7,4	0,6	70,4	1,3	43,7	3,0	1,4	42,0	**1**	**1**	*Correta*
BAIC	21,4	-19,4	0,7	72,3	0,9	328,0	0,5	0,2	59,8	**0**	**0**	*Correta*
	22,8	2,5	0,6	69,6	0,9	100,5	1,9	0,6	67,3	**1**	**0**	*Incorreta*
BMW	24,8	8,1	0,7	77,2	1,0	35,1	4,3	0,7	48,8	**0**	**1**	*Incorreta*
	24,9	8,3	0,8	75,8	1,0	35,3	4,4	0,6	47,6	**1**	**1**	*Correta*
	25,0	1,7	0,8	80,0	1,0	25,2	4,8	0,6	48,6	**0**	**0**	*Correta*
	25,1	8,4	0,9	78,8	1,1	27,5	5,3	0,6	46,8	**1**	**1**	*Correta*
	25,2	11,7	0,8	78,1	1,0	32,6	6,0	0,6	49,0	**1**	**1**	*Correta*
	25,4	10,8	0,8	76,0	1,0	43,7	2,8	0,6	50,2	**1**	**1**	*Correta*
	25,5	10,0	0,7	75,5	0,9	44,6	2,7	0,6	50,9	**1**	**1**	*Correta*
BRILLIANCE	21,2	-8,9	0,9	60,5	1,1	59,2	2,6	0,7	81,5	**0**	**0**	*Correta*
	21,5	-3,1	0,8	63,4	1,1	80,0	3,8	0,9	84,3	**0**	**0**	*Correta*
	20,6	8,9	1,1	31,5	1,0	205,9	3,6	0,4	86,3	**1**	**1**	*Correta*
	20,5	9,3	1,0	27,0	0,9	219,4	3,0	0,3	86,3	**1**	**1**	*Correta*
	20,4	10,0	1,0	28,9	1,1	253,5	2,7	0,2	92,2	**1**	**0**	*Incorreta*
BYD	21,4	20,1	0,5	67,7	0,9	114,6	2,0	0,9	85,7	**1**	**1**	*Correta*
	22,6	7,3	0,5	60,8	0,7	107,1	2,1	1,0	78,3	**1**	**1**	*Correta*
	22,6	5,3	0,5	62,1	0,7	130,8	1,7	0,8	75,7	**0**	**0**	*Correta*
	22,8	4,8	0,6	73,0	0,8	209,5	1,6	0,6	77,2	**0**	**0**	*Correta*
	22,9	4,9	0,5	72,9	0,8	189,8	1,7	0,7	74,4	**0**	**0**	*Correta*
CHANGAN	21,4	5,0	0,7	47,2	1,0	120,3	2,2	0,7	94,2	**0**	**0**	*Correta*
	22,0	4,3	0,8	64,0	1,0	77,5	5,8	1,3	90,3	**1**	**1**	*Correta*
	22,2	3,1	0,6	66,4	0,8	99,8	2,4	0,7	80,1	**0**	**0**	*Correta*
DAIMLER	26,0	1,4	0,8	82,0	1,0	41,1	2,1	0,8	65,7	**0**	**0**	*Correta*
	25,8	8,2	0,8	73,3	1,1	38,7	2,7	0,8	50,6	**1**	**1**	*Correta*
	26,0	8,5	0,9	75,3	1,2	32,1	3,5	0,7	47,1	**1**	**1**	*Correta*
	26,0	8,5	0,9	75,3	1,2	32,1	3,5	0,7	47,1	**1**	**1**	*Correta*
DONGFENG	22,7	4,0	0,5	26,8	0,3	0,7	91,6	3,2	54,9	**1**	**1**	*Correta*
	23,1	10,1	3,1	61,8	1,1	78,8	7,5	1,9	23,5	**1**	**1**	*Correta*
	23,7	10,9	2,0	55,4	1,3	116,8	6,6	1,2	55,7	**1**	**1**	*Correta*
	23,6	10,1	3,0	53,0	1,4	95,9	5,2	1,1	34,5	**1**	**1**	*Correta*
	22,4	4,1	2,6	45,6	1,2	185,3	2,1	0,3	38,2	**1**	**1**	*Correta*
	23,2	6,0	1,7	49,3	1,0	76,7	7,8	0,6	48,5	**1**	**1**	*Correta*
	23,5	5,6	2,4	49,3	1,2	59,4	10,2	0,8	39,9	**1**	**1**	*Correta*
FAW	21,1	4,0	1,4	34,6	2,0	97,8	4,7	1,2	97,9	**0**	**0**	*Correta*
	22,3	0,6	1,1	48,7	1,4	65,8	8,3	1,9	94,1	**0**	**0**	*Correta*
	22,0	-4,2	0,8	53,4	1,1	80,1	5,1	1,4	94,2	**0**	**0**	*Correta*
	22,2	-0,2	0,5	52,1	1,0	93,9	4,4	1,4	93,6	**0**	**0**	*Correta*
FIAT	25,1	5,4	1,5	82,4	1,7	101,9	5,4	1,0	44,7	**1**	**1**	*Correta*
	25,0	5,0	1,5	83,2	1,9	103,3	5,0	1,0	38,8	**1**	**1**	*Correta*
	24,9	0,7	1,8	84,7	2,3	107,8	3,9	0,8	32,6	**0**	**0**	*Correta*
	25,0	2,8	1,5	84,3	1,6	128,6	3,2	0,5	57,0	**0**	**0**	*Correta*
	25,1	5,6	1,1	89,1	1,4	92,7	3,9	0,8	35,6	**0**	**1**	*Incorreta*
	25,6	1,2	1,0	86,6	1,4	81,4	3,9	1,0	36,4	**0**	**0**	*Correta*
	25,7	1,1	1,0	86,3	1,3	76,9	4,2	1,1	38,4	**0**	**0**	*Correta*

EMPRESA	T	MO	LS	PCT	LC	PMPC	GAI	GAT	CE	Grupo *Clustering*	Grupo Predito	Classificação
FORD	25,7	3,7	1,9	86,9	2,0	55,3	5,6	0,8	42,2	**1**	**1**	*Correta*
FUJI	23,4	4,0	0,6	65,5	1,0	65,2	2,7	1,1	71,1	**0**	**0**	*Correta*
	23,4	3,2	0,7	62,5	1,1	64,2	2,7	1,1	68,6	**0**	**0**	*Correta*
	23,4	1,9	0,8	69,1	1,2	57,8	3,1	1,2	65,2	**0**	**0**	*Correta*
	24,1	14,7	1,3	53,5	1,7	54,1	5,9	1,4	74,2	**1**	**1**	*Correta*
GEELY	16,7	-6,8	1,2	47,4	1,2	127,3	12,7	0,1	24,4	**1**	**1**	*Correta*
	16,8	-25,2	10,9	19,7	1,2	114,7	6,1	0,1	13,3	**0**	**1**	*Incorreta*
	21,8	10,7	1,4	60,8	1,3	198,9	4,0	0,9	73,5	**1**	**1**	*Correta*
GM	25,6	3,8	0,9	74,0	1,1	61,8	7,2	1,0	45,9	**1**	**1**	*Correta*
	25,8	3,3	1,1	74,4	1,3	64,8	6,0	1,0	50,4	**0**	**0**	*Correta*
GREATWALL	17,8	22,2	1,8	39,2	2,2	102,9	4,2	0,7	82,1	**1**	**1**	*Correta*
	20,9	7,2	0,8	55,4	1,4	83,5	4,4	0,9	85,0	**1**	**0**	*Incorreta*
	21,3	7,6	0,6	86,5	1,3	80,3	3,6	1,0	85,0	**1**	**1**	*Correta*
	21,9	13,6	1,3	57,9	1,3	84,3	3,8	1,1	73,9	**1**	**1**	*Correta*
	22,2	13,3	0,7	87,9	1,4	88,3	4,1	1,1	85,0	**1**	**1**	*Correta*
	22,8	17,0	1,2	46,8	1,4	87,4	3,5	1,2	92,8	**1**	**1**	*Correta*
	22,9	14,8	1,2	45,5	1,4	100,0	3,0	1,1	93,7	**1**	**1**	*Correta*
HONDA	25,4	7,7	0,9	62,8	1,2	51,8	5,7	1,0	56,8	**1**	**1**	*Correta*
	25,5	7,9	0,9	64,0	1,1	46,0	4,6	1,0	58,0	**1**	**1**	*Correta*
	25,2	6,4	1,1	61,5	1,3	42,0	2,7	0,8	50,1	**1**	**1**	*Correta*
	25,1	2,9	1,0	62,6	1,3	50,4	2,4	0,7	48,5	**0**	**0**	*Correta*
	25,7	4,5	0,9	61,1	1,2	32,5	2,2	0,8	46,2	**0**	**0**	*Correta*
ISUZU	23,5	5,7	0,8	79,1	1,1	78,0	3,4	1,4	53,8	**1**	**1**	*Correta*
	23,5	6,4	1,0	72,2	1,2	78,4	3,5	1,4	54,6	**1**	**1**	*Correta*
	23,4	1,5	0,8	72,7	1,2	69,7	2,8	1,3	45,3	**0**	**0**	*Correta*
	23,4	6,2	1,0	70,5	1,3	71,1	2,9	1,3	52,0	**1**	**1**	*Correta*
	23,5	7,9	1,0	60,6	1,3	81,9	3,4	1,3	62,9	**1**	**1**	*Correta*
	23,6	9,1	1,2	58,1	1,6	75,8	3,2	1,1	55,9	**1**	**1**	*Correta*
MAHINDRA	22,4	7,7	1,8	77,5	2,2	106,2	4,4	1,0	34,6	**1**	**1**	*Correta*
MAZDA	24,1	4,2	0,6	77,8	0,9	51,7	3,5	1,6	60,2	**1**	**1**	*Correta*
	23,8	0,4	1,0	73,9	1,3	51,5	2,5	1,2	47,0	**0**	**0**	*Correta*
	23,9	1,0	1,0	75,8	1,3	50,9	2,9	1,3	47,8	**0**	**0**	*Correta*
	24,1	6,7	1,0	64,8	1,5	64,2	3,4	1,3	56,4	**0**	**0**	*Correta*
	24,2	6,6	1,1	61,5	1,5	48,7	3,6	1,4	59,7	**1**	**1**	*Correta*
MITSUBISH	23,8	0,3	0,7	82,8	1,0	123,4	4,1	1,4	67,3	**0**	**0**	*Correta*
	23,7	0,2	0,6	81,2	0,9	108,7	4,4	1,4	67,1	**0**	**0**	*Correta*
	23,4	1,0	0,6	82,2	0,9	95,9	3,4	1,2	74,8	**0**	**0**	*Correta*
	23,6	2,2	0,8	81,8	1,1	88,0	4,6	1,4	65,2	**0**	**1**	*Incorreta*
	23,6	3,7	0,9	76,6	1,1	103,5	4,8	1,3	70,8	**0**	**1**	*Incorreta*
	23,8	5,9	1,0	65,0	1,3	86,2	5,3	1,4	71,9	**1**	**1**	*Correta*
NISSAN	25,4	7,4	1,0	71,1	1,2	0,0	2,2	0,9	63,2	**1**	**1**	*Correta*
	25,4	7,3	1,0	70,6	1,2	48,6	2,3	0,9	62,2	**1**	**1**	*Correta*
	25,4	4,8	1,4	70,5	1,7	60,2	2,3	0,8	50,1	**0**	**0**	*Correta*
PEUGEOT	25,0	2,0	0,9	80,2	1,0	83,4	3,7	0,8	80,4	**0**	**0**	*Correta*
	25,1	1,9	0,9	79,4	1,1	79,2	4,1	0,9	80,5	**0**	**0**	*Correta*
	24,9	-2,9	1,0	80,8	1,1	75,2	3,5	0,8	73,1	**0**	**0**	*Correta*
	25,1	1,5	0,9	80,0	1,0	70,6	4,3	0,9	76,0	**0**	**0**	*Correta*
	25,0	1,8	0,9	80,3	1,0	73,6	4,8	0,9	69,0	**0**	**0**	*Correta*
RENAULT	24,7	3,7	0,9	71,9	1,0	82,9	3,6	0,6	82,0	**1**	**0**	*Incorreta*
	24,7	2,1	0,9	69,9	1,0	82,7	3,2	0,6	82,9	**1**	**0**	*Incorreta*

EMPRESA	T	MO	LS	PCT	LC	PMPC	GAI	GAT	CE	Grupo *Clustering*	Grupo Predito	Classificação
SAIC	22,2	5,7	1,0	63,4	1,2	107,1	1,7	0,6	75,8	0	0	*Correta*
	24,6	8,6	1,0	71,1	1,1	56,0	12,2	1,7	81,3	1	1	*Correta*
	25,3	6,5	0,9	64,1	1,1	50,2	14,0	1,5	74,3	1	1	*Correta*
SUZUKI	24,2	4,2	0,9	63,1	1,3	82,6	5,6	1,5	77,3	0	0	*Correta*
	23,9	3,2	1,3	60,0	1,6	73,8	4,3	1,1	65,3	0	0	*Correta*
	23,9	4,8	1,2	57,1	1,5	59,3	4,9	1,1	78,9	0	0	*Correta*
	24,0	5,6	1,5	53,9	1,8	67,1	4,6	1,1	66,0	0	0	*Correta*
	24,1	6,4	1,4	53,8	1,7	66,8	4,5	1,1	68,3	0	0	*Correta*
TATA	22,6	8,0	0,7	66,3	0,9	52,5	6,1	1,4	67,2	1	1	*Correta*
	22,7	6,5	0,5	71,5	0,7	63,8	4,5	1,1	69,7	1	1	*Correta*
	23,4	-4,2	0,3	95,6	0,5	75,3	4,5	1,3	77,8	0	0	*Correta*
	23,9	7,1	0,5	79,8	0,8	89,5	5,4	1,3	66,7	1	1	*Correta*
	24,2	4,7	0,7	77,0	0,9	92,8	6,5	1,4	65,8	1	1	*Correta*
TOYOTA	26,1	8,9	0,9	63,2	1,1	42,4	3,3	0,8	55,2	1	1	*Correta*
	26,0	-2,3	0,9	65,4	1,1	34,6	2,7	0,7	55,7	0	0	*Correta*
	26,0	2,5	1,0	65,4	1,1	38,0	2,9	0,6	55,4	0	0	*Correta*
	26,1	6,0	0,9	65,8	1,1	42,7	3,4	0,7	55,3	0	0	*Correta*
VOLKSWAGEN	26,2	6,0	0,8	75,0	1,1	38,9	4,2	0,7	45,5	1	1	*Correta*
	26,3	6,3	0,8	74,4	1,0	41,3	4,6	0,6	50,0	0	1	*Incorreta*
	26,4	3,1	0,8	74,9	1,0	40,5	3,5	0,6	51,2	0	0	*Correta*
VOLVO	24,3	7,9	0,9	66,4	1,3	67,7	4,7	1,0	61,8	1	1	*Correta*
	24,5	0,0	0,8	74,7	1,1	80,9	4,4	0,9	56,1	0	0	*Correta*
	24,4	2,6	0,8	78,0	1,1	87,0	3,4	0,8	56,4	0	0	*Correta*

As variáveis preditoras estão representadas sob a seguinte forma:

- T = Tamanho;
- MO = Margem Operacional;
- LS = Liquidez Seca;
- PCT = Participação de Capitais de Terceiros;
- LC = Liquidez Corrente;
- PMPC = Prazo Médio de Pagamento de Compras;
- GAI = Giro do Ativo Imobilizado;
- GAT = Giro do Ativo Total;
- CE = Composição do Endividamento.

TAXA DE ACERTOS:

Grupo I	Grupo II	TOTAL
89,3%	**91,9%**	**90,7%**