UNIVERSIDADE PEDAGÓGICA DE MAPUTO

Faculdade de Ciências Naturais e Matemática

Curso de Mestrado em Estatística

FACTORES ASSOCIADOS AOS NÍVEIS DE SATISFAÇÃO E OPÇÕES DA CLIENTELA DOS SUPERMERCADOS DE MAPUTO

Autor:

Salomão Eugénio Munguambe

Supervisora:

Professora Doutora Aida Calviño Martinez

Maputo, Julho de 2024

Salomão Eugénio Munguambe

FACTORES ASSOCIADOS AOS NÍVEIS DE SATISFAÇÃO E OPÇÕES DA CLIENTELA DOS SUPERMERCADOS DE MAPUTO

Dissertação submetido ao Departamento de Estatística
da Faculdade de Ciências Naturais e Matemática,
para a obtenção do grau de Mestre.

Supervisora:

Professora Doutora Aida Calviño Martinez

Maputo, Julho de 2024

Índice

Lista de Tabelas

Lista de Figuras

Siglas

AIC – Akaike information criterion

ARI – Adjasted Rand Index (Índice de Rand Ajustado)

IOF – Inquérito sobre o Orçamento Familiar

MRLB – Modelo de Regressão Logística Binária

MRLO – Modelo de Regressão Logística Ordinal

MDS – Multidimentional Scaling (Escalonamento Multidimensional)

STRESS – Standardized Residual Sum of Squares

Declaração

Declaro que esta dissertação é resultado da minha investigação pessoal e das orientações da minha supervisora. O seu conteúdo é original e todas as fontes consultadas estão devidamente mencionadas no texto e nas referências bibliográficas.

Declaro ainda que este trabalho, não foi apresentado em nenhuma outra instituição para a obtenção de qualquer grau académico.

Maputo, Julho de 2024

_____

(Salomão Eugénio Munguambe)

Dedicatória

A dedicação vai para a minha família, especialmente, a minha esposa Catiça Adamp Halde Munguambe pelo apoio que me prestou e pela sua paciência por não ter sido disponível em alguns momentos em que era suposto estar. Aos meus filhos, Pablo, Karina Amélia e Geovanna que em varios momentos foram privados do calor e amor de pai, mas sempre demenstraram estar a perceber que algo de importante no momento tomava a atenção do pai, mas que em breve faria parte do passado.

Agradecimentos

Em primeiro lugar agradeço a Deus todo poderoso pelo dom da vida, pela seúde, por nunca ter permitido que me faltassem forças nesta caminhada académica.
Agradecimento muito especial à Professora Doutora Aida Calviño Martinez pela disponibilidade, rigor cientifico e compreensão na supervisão deste trabalho.

Aos meus irmãos pela confiança e encorrajamento que me transmitiram e que me serviu de combustível sempre que o desânimo quesesse tomar conta de mim.
Aos colegas da edição 2020 – 2022 do curso do Mestrado em Estatística, por ter contribuido não só para minha formação científica mas também humana.

E mais importante ainda, a todos professores desta edição que deram o seu melhor para nos formar de modo a sermos actores válidos na construção do nosso país, e buscaram sempre transmitir uma extraordinária experiência de amizade entre os povos.

Resumo

O presente trabalho com o tema *Factores Associados aos Níveis de Satisfação e Opções da Clientela dos Supermercados de Maputo,* procura dar respostas aos consumidores e fornecedores de serviços de supermercados e não só, da região metropolitana de grande Maputo e província de Maputo.

Procura orientar tanto os consumidores em termos de opções de locais de compras que melhor respondem as suas expectativas, assim como aos investidores ou prestadores destes mesmos serviços no que concerne às características do público que pretende alcançar ou servir.

Para dar resposta a estas pretenções recorre-se aos métodos estatísticos de análise multivariada de dados, designadamente, *Regressão Logística Binária* para determinar a probabilidade de um cliente ou potencial cliente com certas caracteráisticas, optar pelo supermercado *Game* ou *Shoprite,* pretende-se tambem com recurso à árvore de classificação, determinar as características de clientes que são favoráveis aos supermercados *Game* ou *Shoprite.*

Pretende-se também com recurso à técnica multivariada *Regressão Logística Ordinal,* determinar a probabilidade de um cliente ou potencial cliente com certas caracteríscas, alcançar certo nível de satisfação e, em adição a esta técnica, vamos construir uma *Árvore de Classificação Ordinal* que nos permitirá classificar os clientes que alcansam os diferentes níveis de satisfação.

E por fim, de modo a tirar maior proveito da extensão da base de dados em uso neste estudo, a título exploratório, vamos aplicar a técnica multivariada *Escalonamento Multidimensional (MDS),* para estudar as similaridades existentes entre as variáveis que não forem tidas como releventes pela litatura, e agrupá-las conforme as suas similaridades de modo que, estudos posteriores possam assumí-las como tal.

Palavras-chave: Satisfação dos clientes, Modelos de regressão logística binária e Ordinal, Supermercados de Maputo.

Abstract

This Research, with the theme *Factors Associated with Levels of Satisfaction and Customer Options in Maputo Supermarkets*, seeks to provide answers to consumers and service providers in supermarkets and beyond, in the greater Maputo metropolitan region and Maputo province.

It seeks to guide both consumers in terms of shopping options that best meet their expectations, as well as investors or providers of these services with regard to the characteristics of the public it intends to reach or serve.

To respond to these intentions, statistical methods of multivariate data analysis are used, namely, *Binary Logistic Regression* to determine the probability of a customer or potential customer with certain characteristics, opting for *Game* or *Shoprite* supermarket. To the *classification tree*, determine the characteristics of customers who are favorable to Game or Shoprite supermarkets.

It is also intended, using the multivariate *Ordinal Logistic Regression* technique, to determine the probability of a customer or potential customer with certain characteristics, achieving a certain level of satisfaction and, in addition to this technique; we will build an *Ordinal Classification Tree* that will allow us to classify customers who achieve different levels of satisfaction.

And finally, in order to take greater advantage of the extension of the database in use in this study, on an exploratory basis, we will apply the multivariate technique *Multidimensional Scaling* (MDS), to study the similarities between variables that are not considered relevant by litature, and group them according to their similarities so that later studies can assume them as such.

Keywords: Customer satisfaction, Binary and Ordinal logistic regression models, Supermarkets

Maputo.

1 Capítulo I

1.1 Introdução

1.1.1 Contextualização

A satisfação dos clientes constitui actualmente uma das maiores prioridades dos gestores de empresas de fornecimento de bens e serviços. Deste modo as empresas empenham-se no sentido de alcançar elevados padrões de qualidade dos seus produtos e serviços, e maximizar os resultados alcançados junto dos seus clientes (KOTLER & KELLER, 2006).

De outro lado, temos os consumidores destes bens e serviços que procuram satisfazer as suas expectativas junto aos fornecedores ou provedores dos mesmos, sendo a satisfação destas expectativas dependente das suas concepções, origens, condições sociais, sexo, idade, etc. Sendo a satisfação uma medida de escala ordinal, a presente investigação pretende encontrar um modelo de regressão logística ordinal capaz de prever a probabilidade de um cliente alcançar um certo nível de satisfação. Também se construirá um modelo de regressão logística binária capaz de prever a probabilidade de um cliente escolher um supermercado entre as duas maiores cadeias de supermercados da cidade e província de Maputo, G*ame e Shoprite*. Pretende-se contribuir não só na resposta das expectativas dos consumidores, mas também ajudar no mapeamento das zonas urbanas mais receptivas a certos tipos de infra-estruturas comercias ou qualidades de serviços.

1.2 Delimitação do Tema

Segundo Gil (2008, pag. 17), "...A delimitação refere-se ao planeamento da pesquisa em sua dimensão mais ampla, envolvendo tanto a sua diagramação quanto a previsão de análise e interpretação dos dados. Entre outros aspectos, a delimitação considera o ambiente em que são coletados os dados, bem como as formas de controle das variáveis envolvidas".

No presente trabalho pretendemos estudar a satisfação dos clientes da cidade e provincia de Maputo, as probabilidades destes clientes optarem pelo supermecado Game ou Shoprite como supermercado de sua preferência e, queremos através da técnica de Escalonamento Multidimensional (MDS), representar agrupamentos das variáveis (características) de clientes, conforme as suas semelhanças ou similaridades. A partir de uma base de dados produzida num inquérito realizado sobre os clientes e potenciais clientes dos grandes supermercados da cidade e provincia de Maputo, com maior enfoque às duas grandes redes de supermercados, *Game* e *Shoprite*, um estudo realizado em agosto de 2013, pela ONG Direito do Consumidor (vede o questionario do inquérito em anexo).

1.3 Justificativa

Nos dias que correm e com a expansão da área urbana, é comum encontrar grandes empreendimentos comerciais que foram construídos e abertos mas em tão curto espaço de tempo fecharam ou foram reorientados a outros fins por se manifestar economicamente inviáveis, a título de exemplo podemos nos referir às actuais instalações do actual ISCIM – Instituto Superior de Comunicação e Imagem, que foram construidas para servir como um pequeno centro comercial (*Mall),* e chegou a funcionar um banco e uma loja de telefonia móvel, mas por estar localizado num bairro densamente habitado por pessoas de baixa renda e perto do maior mercado informal – Xipamanine, que vende todo tipo de artigos desde vestuário até aos alimentares passando pelo material de construção, teve que ser reorientando para outro fim. Exemplos similares são enúmeros. E porque para que um empreendimento seja economicamente viável, dentre vários elementos que devem constar no plano de negócio, fundamentalmente deve existir um mercado consumidor – os clientes. O presente trabalho pretende criar modelos de *regressão logística binária* e *ordinal* capazes de prever a probabilidade de uma pessoa (consumidor) optar por um ou outro supermercado (entre *Game e Shoprite)* e alcançar os diferentes níveis de satisfação, respectivamente, tendo em conta as suas diferentes características, nomeadamente: sexo, idade, faixa salarial, escolaridade, etc. Pretende-se com recurso às técnicas estatísticas criar modelos com boa capacidade de

previsão (eficiência e especificidade), que possam auxiliar os investidores nas áreas de vendas tais como supermercados, restaurantes e outros serviços prestados na região de grande Maputo como espaço geográfico de estudo, e não só.

1.4 Perguntas de Pesquisa

Face ao que foi exposto acima em termos de justificativa, mostra-se que a nossa pesquisa pretende responder as seguintes questões:

I. Qual é o perfil do cliente que é favorável ao supermercado *Game ou Shoprite?*
II. Qual é o perfil do cliente que alcança um determinado nível de satisfação junto aos supermercados da província e cidade Maputo?
III. Quais são os grupos de variáveis ou características dos clientes que apresentam similaridades?

1.5 Objectivos

a) Objectivos Gerais

- Auxiliar os investidores nas áreas de supermercados ou vendas no geral e os respectivos clientes ou potenciais clientes, sobre as melhores opções em termos de áreas mais receptíveis a certo tipo de serviços e melhores opções para compras, respectivamente.

b) Objectivos Específicos

- Criar modelos de regressão logística binária e ordinal capazes de prever a probabilidade de um cliente optar por um supermercado e, alcançar os diferentes níveis de satisfação;
- Criar árvores de classificação capazes de classificar os clientes em termos de opções de supermercados e o alcance de diferentes níveis de satisfação;
- Encontrar as similaridades existentes nos atributos das observações.

2 Capítulo II

2.1 Revisão Bibliográfica

Segundo Kotler e Keller (2012), no capítulo "O que influencia no comportamento do consumidor?", em resposta, lista os factores culturais, sociais e pessoais, psicológicos (motivacionais, percepção, aprendizagem, emoções e memória), são fundamentais na decisão sobre onde obter serviços ou bens de que necessita.

Para Bochado, A. O., Caetano, J., Cobra, J., etc.(2013) "... os consumidores procuram via da regra geral manter um comportamento de compras de acordo com o que percebem ser os valores do seu grupo de referência – grupus cujas estruturas de valores influenciam o comportamento de uma pessoa."

Num outro dezenvolvimento, os autores apresentam-nos um outro conceito não menos importante que é o conceito de *segmentação do mercado.*

"segmentar um mercado corresponde na prática a dividí-lo em conjunto de consumidores mais ou menos semelhantes mas também é um processo de escolha. A empresa escolhe que clientes vai servir e que clientes não vai servir, que mercados vai abranger e que mercado não vai abranger.

Na prática de segmentação o acto de definir o que não vai se fazer, é muitas vezes tanto ou mais importante que definir o que vai se fazer".

Segundo Shaw & John (2002), no capítulo "*tsunami das experiências do cliente",* enfatiza a experiência emocional do cliente como uma diferenciadora chave nos tempos modernos, dada a crescente indiferenciação dos produtos por todos mercados causada pela difusão da internet. Sendo assim, os diferenciadores tradicionais como o preço, distribuição, rapidez do processo, etc, deixaram de ser estratégias de negócios sustentáveis.

Fauque J. (1993), apresenta-nos o atendimento como uma marca de excelência, essa marca de excelência possibilita estabelecer a diferença diante da concorrência, cria imagem da marca, dá provas da posição resolutamente favorável ao consumidor cliente.

Moreira (2010), no capítulo IX, "foco *no cliente",* diz que o cliente será o grande juiz quanto à qualidade do atendimento, que poderá ir do superior, a média ou inferior, dependendo da experiência e do que o cliente entende por qualidade.

Fundamenta ainda que*:*

“ A avaliação do cliente irá incidir: no grau de prestação de serviços, na medida em que este satisfaz as suas exigências, pretensões ou expectativas, atitudes do profissional do atendimento, respostas dadas em tempo útil, cumprimento de processos, capacidade de assumir responsabilidades, as formas como as reclamações são atendidas, etc”.

Vilares M. J. & Coelho P. S. (2011,P. 27) definem satisfação em duas vertentes a saber:

Satisfação como transação específica de cliente e Satisfação como um processo cumulativo. Na primeira vertente os autores entendem a satisfação como um juízo avaliativo pós-escolha relativo a uma compra ou transação específica. Na segunda vertente, entendem satisfação como uma avaliação global baseada no conjunto de experiências de compra e consumo de produtos ou serviços da empresa ao longo do tempo.

Num outro conceito, mas relacionado, ainda que diferente, os autores apresentam-nos a lealdade de clientes. Mas antes da discussão deste conceito advertem-nos sobre a necessidade de saber distinguir a lealdade da retenção. Citando Johnson e Gustasson (2000, p 7), definem a lealdade como intenção ou pré disposição do cliente em comprar de novo. Enquanto a retenção de clientes pode dever-se a outros factores como preços praticados, proximidade do fornecedor e principalmente a existência de um único fornecedor.

Neste desenvolvimento apresentam-nos a lealdade em três dimensões a saber:

Lealdade afectiva – tem ligações emocionais e envolve interacções humanas.

Lealdade racional ou cognitiva – envolve avaliações de relações comerciais, incluindo avaliações de preços, valor recebido, custos, ets.

Lealdade comportamental – resulta das duas dimensões anteriores e corresponde à intensão de continuar cliente e de recomendar a empresa.

Enquanto que, para Bochado. *at all* (2013), “...a satisfação é uma manifestação de agrado de cliente como resultado do desempenho do producto face às suas expectativas, contribuindo para situação de permanência e recompra ou afastamento e litigância”.

2.2 Estudos similares em alguns países

Agresti, A. (2007, p. 182), apresenta um estudo sobre uma Pesquisa Social Geral, que relaciona a ideologia política com o partido político a que se está filiado e o género. A ideologia política tem uma escala ordinal de cinco categorias, variando de muito liberal a muito conservador. Ele toma como variáveis preditoras o género e o partido político a que se está filiado. Para responder a questão da sua pesquisa recorre ao modelo de regressão logística ordinal e realiza os testes necessários para verificar a robustez do modelo. No presente trabalho, com a base em análise queremos executar os procedimentos similares a estes para podermos apresentar um modelo de regressão logística ordinal que se manifeste aceitável.

João I., em publicação na revista "*Turismo e Desenvolvimento - 17/18 de 2012"* **,** páginas *87 – 96,* apresenta o trabalho intitulado: *"Aplicação da regressão logística ordinal em estudos de lealdade de clientes. Evidência para a indústria hoteleira no Algarve",* onde a autora apresenta um estudo sobre a lealdade dos clientes (turistas) de hoteis que tomam a região de Algarve, em Portugal como seu destino turísco. o estudo toma como variável resposta a lealdade dos clientes com a escala ordinal (1= "decerto recomenda", 2 = "provavelmente sim", 3 = "talvez sim/talvez não", 4 = "provavelmente não" e 5 = "decerto não"), isto é, em que medida estes recomendariam ou não o hotel onde se hospedaram em função das variáveis regressoras: *idade, tempo de estadia, etc.* Para responder esta questão a autora recore à metodologia estatística dos modelos de *odds proporcionais* e *odds proporcionais parciais*.

No presente trabalho queremos estudar outra vertente de um estudo similar, que é na verdadeda o que motiva a recomendação dos servicos, a satisfação daquele que usufruiu dos serviços prestados. Nesta vertente os serviços em causa são de compras em supermercados. Neste caso, queremos estudar de que depende a satisfação da clientela dos supermercados da região de grande Maputo.

Um estudo similar que serve de referencia para o nosso estudo é a Monografia Científica do estudante **Leandro José Calsing,** do Instituto Universitário UNIVATES – Centro de Gestão Organizacional - Curso de Gestão Organizacional, Intitulado "*Estudo sobre a satisfação do clientes da Fábrica de Móveis Klein Ltda",* cujo objectivo é estudar a satisfação dos clientes da empresa de fábrica de móveis klein Ltda, com a escala ordinal com cinco categorias que varia de *1="totalmente insatisfeito"* a *5 = " totalmente satisfeito"*. O estudo é realizado sobre os clientes da empresa entre os anos 2005 a 2007, procurando identificar o perfil do cliente da empresa, o atributo que leva o cliente a comprar os productos da empresa, o factor de maior e menor satisfação, nível de satisfação com o producto, atendimento e entrega. Para a prossecução do estudo o autor recorre ao método censatório (censo) da

população, uma vez que o estudo é feito numa pequena empresa e é feito o censo na carteira de clientes num período de dois anos, para o tratamento dos dados desse censo recorre-se à estatística descritiva.

No nosso estudo, além de explicar o alcance dos diferentes níveis de satisfação, a probabilidade de potenciais clientes vierem a optar por supermercado Game ou Shoprite, queremos tambem agrupar as variáveis que caracterizam os clientes ou potenciais clientes, coforme estas se manifestem ser ou não semelhantes.

Ricardo Fasti de Souza e Wilton de Oliveira Bussab no seu artigo " ***Participação de Mercado: Uma Abordagem Multidimensional Escalar***", apresentam um estudo do mercado que visa mapear as tendências de preferências de consumidores em relação aos produtos e marcas entre os consumidores. O estudo visa analisar a relação entre participação de mercado e EMD na tentativa de oferecer ao gestor de programas de marketing uma ferramenta que antecipe tendências em relação às preferências e seus impactos sobre o consumo e consequentemente sobre a participação de mercado.

2.3 Abordagem da Pesquisa

Segundo Gil (2008, pag. 17) a aplicação da teoria estatística da probabilidade constitui importante auxílio para a investigação em ciências sociais. Há que se considerar, porém, que as explicações obtidas mediante a utilização do método estatístico não podem ser consideradas absolutamente verdadeiras, mas dotadas de boa probabilidade de serem verdadeiras.

Mediante a utilização de testes estatísticos, torna-se possível determinar, em termos numéricos, a probabilidade de certa determinada conclusão, bem como a margem de erro de um valor obtido. Portanto, o método estatístico passa a caracterizar-se por razoável grau de precisão, o que o torna bastante aceite por parte dos pesquisadores com preocupações de ordem quantitativa.

Os procedimentos estatísticos fornecem considerável reforço às conclusões obtidas, sobretudo mediante a experimentação e a observação. Tanto é que os conhecimentos obtidos em alguns setores da Psicologia e da economia devem-se fundamentalmente à utilização do método estatístico.

Gil (1999) menciona que a pesquisa qualitativa é subjectiva ao objecto de estudo, ergue-se sobre a dinâmica e abordagem do problema pesquisado e visa descrever e descodificar de forma interpretativa os componentes de um sistema complexo. Pesquisar qualitativamente é analisar, observar, descrever e realizar práticas interpretativas de um fenômeno a fim de compreender seu significado.

Nesta perspectiva, quanto à abordagem a nossa pesquisa é quantitativa com alguma componente qualitativa (mista). Com base nas perguntas de pesquisa, esperamos recorrer a análises não necessariamente quantitativas em toda sua extenção. esperamos que, com base em estudos comportamentais e outras literaturas expliquemos a natureza das respostas das perguntas de pesquisa.

2.4 Procedimentos de Colecta de Dados

Segundo Mattar (1996, pag. 48), "... dados secundarios são aqueles que foram colectados, tabelados, ordenados e, às vezes, até analizados e catalogados e que são colocados à disposição dos intereçados. As fontes básicas dos dados secundários são: a própria empresa, governos, publicações e organizações não governamentais...".

Para Kottler (2004), "... os dados secundários são informações que já existem em algum lugar e que foram colectados para outro propósito".

A presente pesquisa será realizada a partir de dados secundários cuja base de dados foi produzida num inquérito realizado sobre os clientes e potenciais clientes dos grandes supermercados da cidade e provincia de Maputo, com maior enfoque às duas grandes redes de supermercados, *Game* e *Shoprite,* um estudo realizado em agosto de 2013, pela ONG Direito do Consumidor, onde foram inqueridas 386 pessoas (vede o questionario do inquérito em anexo).

2.5 Análise de Dados

A análise dos dados é uma das fases mais importantes da pesquisa, pois, a partir dela, é que serão apresentados os resultados e a conclusão da pesquisa, conclusão essa que poderá ser final ou apenas parcial, deixando margem para pesquisas posteriores (MARCONI & LAKATOS, 1996).

Há diversas técnicas de análise de dados que podem ser utilizadas em pesquisas de natureza qualitativa ou quantitativa. De acordo com Trivinõs (1987, p. 137) , é possível concluir que todos os meios que se usam na investigação quantitativa podem ser empregados também no enfoque qualitativo. Sendo assim, o que varia é o enfoque, isto é, a atenção especial ao informante, ao mesmo observador e às anotações de campo, o que não ocorre na pesquisa quantitativa.

Enfim, existem várias técnicas de análise de dados, mas as principais são a análise de conteúdo, a estatística descritiva univariada e a estatística multivariada.

Com esta base, e com o auxílio do software estatístico R 4.1.0, MS Excel®2010, IBM SPSS® 20, far-se-á uma partição de 80% para criação (treinamento) dos modelos de *Regressão Logística Binária*, *Regressão Logística Ordinal* e respectivas Árvores de Classificação, e os restantes 20% para testar ou validar o modelo, far-se-há tambem agrupamento de variáveis conforme as suas similaridades, com auxílio da técnica multivariada Escalonamento Multidimensional – MDS.

Resumidamente, para esta pesquisa, vamos executar os seguintes procedimentos estatísticos:

- Descrição dos dados – estatística descritiva;
- Estimação dos modelos de regressão logística binária e ordinal;
- Significância estatística geral dos modelos e dos parâmetros da regressão logística binária e ordinal;
- Efectuar a análise de sensibilidade, eficiência global dos modelos, sensitividade e especificidade;
- Construção e análises de árvores de classificação para os modelos de regressão logística binária e ordinal;
- Aplicar a técica de Escalonamento Multidimnsional – *MDS* para apurar as semelhanças entre os atributos (variáveis) de dados que não entraram na criação dos modelos.

2.6 Variáveis a usar nos modelos propostos

Antes de indicar as variáveis que iremos usar na criação dos modelos dividimos o conjunto das 35 variáveis com 386 observações, cujo formulário do inquérito se encontra em anexo, em 7 grupos conforme a natureza dos seus valores, na tabela resumo a seguir:

Tabela 1: Resumo das variáveis do estudo

Domínio (grupo de variáveis)	**Variável – Discrição**	**Valores Tomados**
Dados ou características do indivíduo	Sexo – Sexo	Feminino – 0 Masculino - 1
	F_etaria – Faixa Etária	Menos de 15 anos – 1 Entre 15 a 25 anos – 2 Entre 26 ba 30 anos – 3 Entre 31 a 40 anos – 4 Entre 41 a 50 anos – 5 Mais de 50 anos – 6
	E_civil – Estado Civil	Soteiro - 1 Casado - 2 Divorsiado – 3 Viuvo – 4
	N_escolar – Nível de Escolaridade	Primario – 1 Basico – 2 Medio – 3 Superior – 4 Nenhum – 5
	F_salarial – Rendimento Mensal	Até 500 MT – 1 Entre 500 a 1500 Mtn – 2 Entre 1500 a 3000 Mtn – 3 Entre 3000 a 4500 Mtn– 4

		Entre 4500 a 10000 Mtn – 5 Entre 10000 a 15000 Mtn – 6 Mais de 15000 Mtn
Informação geral sobre o serviço do supermercado	Sup_preferido – Nome do Supermercado de sua preferência	Game – 1 Shoprite – 2
	Tempo_cliente – A quanto tempo e cliente	Menos de um ano – 1 Entre um e dois anos – 2 Entre 2 e três anos – 3 Mais de 3 anos – 4
	Freq_sup – Com que frequência vem a este Supermercado	Diariamente – 1 Semanalmente – 2 Quinzenalmente – 3 Mensalmente – 4 Ocasionalmente – 5
	Outro_sup – Qual é o outro Supermercado que costuma frequentar	Novo mundo – 1 Ganha pouco – 2 Mohmed e Companhia – 3 LM – 4 Luz – 5 Nosso Supermercado – 6 Outro – 7
	Importancia_sup – Importância do Supermercado	Não é importante – 1 Pouco importante – 2

		Indiferente – 3 Pouco importante – 4 Importante – 5 Muito importante - 6
Tangíveis (Beleza das instalações e	Variedade_serviços– O Supermercado possui vários tipos de serviço	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
forma de apresentação dos funcionários)	Org_productos – Como encontram-se apresentados os produtos	Bem organizados – 1 Organizados – 2 Nem organizados nem desorganizados – 3 Desorganizdos – 4 Muito desorganizados – 5
	Aprumo_func – A apresentação dos funcionários e excelente	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	Acesso_sup – Os locais dos Supermercados são de fácil acesso	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
Confiabilidade (Confiança no seu supermercado, fazer as coisas como prometem,	Cumpromentimento_sup – O Supermercado cumpre sempre o que promete	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5

Satisfação dos serviços prestados, maior qualidade nos produtos e bons preços)	Motivo_escolha – O que mais me faz comprar neste Supermercado	Porque tem todos productos q necessito- 1 Promoções frequentes – 2 A qualidade dos productos é boa – 3 Os preços são baixos – 4 Outros – 5
	MaisValiaNaEscolha – Na hora da escolha do Supermercado o que leva em conta	Atendimento – 1 Ausência de filas – 2 Preço – 3 Variedades – 4 Promoções/ofertas – 5 Qualidade – 6 Localização – 7
	Nível_satisfação – Sente-se satisfeito com este Supermercado	Pouco satisfeito – 1 Pouco – 2 Mais ou menos – 3 Muito – 4 Muito satisfeito – 5
	Melhoria_qualidade – A qualidade dos produtos é melhor que nos outros Supermercados	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	Preço_acessivel – O preço praticado neste Supermercado é mais barato que nos outros Supermercados	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4

		Concordo totalmente - 5
	Empenho_sup – O Supermercado empenha-se em facilitar a vida das pessoas	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
Presteza (Atendimento aos clientes com boa vontade)	Ráp_atendimento – É rápido o atendimento dos caixas no Supermercado	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	Empenho_func – Os funcionários mostram-se sempre disponíveis para o atendimento ao cliente	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	NumStaff_adequado – É adequado o número de funcionários para o atendimento ao cliente	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	Melhores_serv – O Supermercado tem-se esforçado em melhorar a sua qualidade para fornecer melhores serviços	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
Garantia / Segurança (Conhecimento dos funcionários sobre os serviços	Func_compto – Os funcionários transmitem segurança, confiança aos seus clientes pelo seu comportamento	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4

)		Concordo totalmente - 5
	Func_competentes – Os funcionários tem competência e conhecimento suficiente para responder às necessidades do cliente	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
Empatía (Consideração e atenção individualizada ao cliente)	Func_atenc – Os funcionários do supermercado são atenciosos	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	At_perszado – Os funcionários oferecem um atendimento personalizado aos seus clientes	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
	Func_prestativos – Os funcionários resolvem as preocupações e necessidades especificam sempre que entro em contacto com eles	Discordo totalmente – 1 Discordo – 2 Mais ou menos – 3 Concordo – 4 Concordo totalmente - 5
Ordem de valores (a que nível o indivíduo valoriza este atributo)	Tangíveis – Beleza das instalações e forma de apresentação dos funcionários	Mais importante – 1 Importante – 2 Mais ou menos – 3 Não importante - 4 Menos importante – 5
	Confiabilidade – Confiança, fazer coisas que	Mais importante – 1

	prometem, satisfação nos serviços prestados, maior qualidade e bons preços	Importante – 2 Mais ou menos – 3 Não importante - 4 Menos importante – 5
	Presteza – Atendimento ao cliente com boa vontade	Mais importante – 1 Importante – 2 Mais ou menos – 3 Não importante - 4 Menos importante – 5
	Segurança – Conhecimento dos funcionários sobre os serviços	Mais importante – 1 Importante – 2 Mais ou menos – 3 Não importante - 4 Menos importante – 5
	Empatia – Consideração e atenção individualizada ao cliente	Mais importante – 1 Importante – 2 Mais ou menos – 3 Não importante - 4 Menos importante – 5

Fonte:Autor

I. ***Modelo de Regressão Logística Binária***

- **Variável resposta (dependente**) – Sup_preferido (Nome do Supermercado de sua preferência);
- **Variáveis independentes (regressoras**) – *grupo de variáveis ou características do indivíduo* (Sexo, F_etaria, E_civil, N_escolar e F_salarial) e *grupo de variáveis de ordem de valores* (Tangíveis, Confiabilidade, Presteza, Segurançae e Empatia)

II. ***Modelo de Regressão Logística Ordinal***

- **Variável resposta (dependente)** – Nível_satisfação (Nível de satisfação);
- **Variáveis independentes (regressoras**) – *grupo de variáveis ou características do indivíduo* (Sexo, F_etaria, E_civil, N_escolar, F_salarial), *grupo de variáveis sobre Informação geral sobre os serviços do supermercado* **(**Sup_preferido, Tempo_cliente, Freq_sup, Outro_sup e Importancia_sup) e *grupo de variáveis de ordem de valores* (Tangíveis, Confiabilidade, Presteza, Segurança e Empatia).

III. Escalonamento Mutidimensional (MDS)

Para o escalonamento multidimensional, usamos todas variáveis que não foram usadas para a criação dos modelos, a saber:

Org_productos, Aprumo_func, Acesso_sup, Compromentimento_sup, Melhoria_qualidade Preco_acessivel, Empenho_sup, Rap_atendimento, Empenho_func, NumStaff_adequado, Melhores_serv, Func_compto, Func_competentes, Func_atenc, At_perszado e Func_prestativos.

2.6.1 Razões de inclusão e exclusão de variáveis nos modelos

a) Para o modelo de regressão logística binária

I. *Razões de inclusão de variáveis*

Segundo Kotler e Keller (2012), o que influencia na decisão sobre onde obter os serviços ou bens de que o indivíduo necessita, fundamentalmente são factores sociais, pessoais e psicológicos. As variáveis que se enquadram nesta perspectiva são as que caracterizam o cliente (características do cliente) e as que expressam *ordem de valores*, isto é, até que ponto o indivíduo ou cliente valoriza os atributos: Tangíveis, Confiabilidade, Presteza, Segurança e Empatia.

Note que neste parágrafo nos referimos a estas variáveis: Tangíveis, Confiabilidade, Presteza, Segurança e Empatia, como atributos que o indivíduo ou cliente valoriza mais ou menos – são características psicilógicas do indivíduo.

II. *Razões de exclusão de variáveis*

O grupo de variáveis referentes à *Informação geral sobre os serviços do supermercado,* Tangíveis, Confiabilidade, Presteza, Segurança e Empatia são características dos supermercados e não são do individuo ou futuro cliente cuja escolha queremos prever. Logo, não podem ser usados para prever a sua decisão.

Note que neste parágrafo nos referimos a Tangíveis, Confiabilidade, Presteza, Segurança e Empatia como grupo de variáveis que caracterizam os supermercados, não são variáveis.

b) Para o modelo de regressão logística ordinal

I. Razões de inclusão de variáveis

A satisfação é um juízo avaliativo pós-escolha relativo a uma compra ou transacção específica, segundo Vilares M. J. & Coelho P. S. (2011,P. 27), Sendo a satisfação um factor psicológico, que faz parte das características do indivíduo, Segundo Kotler e Keller (2012). Temos o entendimento de que as variáveis que caracterizam o indivíduo como: Sexo, F_etaria, E_civil, N_escolar, F_salarial, Tangíveis, Confiabilidade, Presteza, Segurança e Empatia, podem influenciar na variável resposta que é Nível_satisfação (Nível de satisfação).

Segundo Shaw & John (2002), a experiência de compras ou aquisição de bens ou serviços anterior, que o cliente possui, é fundamental para que este alcance certo nível de satisfação ou frustração ao adquirir produtos ou serviços. Entendemos que, com base neste outor, uma nova experiência para que se transforme numa satisfação ou frustração precisará de uma experiência de compra ou aquisição de serviços anterior a esta que sirva de referência. Sendo assim, torna-se relevante incluir variáveis que trazem respostas para experiência do indivíduo, estamos a falar de variáveis de grupo de variáveis sobre *Informação geral sobre os serviços do supermercado* **(**Sup_preferido, Tempo_cliente, Freq_sup, Outro_sup e Importancia_sup).

II. Razões de exclusão de variáveis

As variáveis referentes aos grupos: Tangíveis, Confiabilidade, Presteza, Segurança e Empatia, dos supermercados abordados no inquérito, são características cujas respostas justificam a satisfação ou

insatisfação que já foi alcançada. Logo, não podem ser usadas para prever níveis de satisfação que elas vem justificar.

c) Para Escalonamento Multidimensional

I. *Razões de inclusão de variáveis*

A razão da escolha das variáveis é devida ao facto destas variáveis não terem encontrado suporte teórico para inclusão das mesmas em modelos de *Regressão Logística Binária* assim como *Ordinal*, e como forma colher informações adicionais que não tenham passado do crivo dos modelos acima citados, e dada a natureza destas variáveis, isto é, elas fazem parte de grupos cujo nível de valor é questionado ao cliente ou potencial cliente dos supermercados *Game* ou *Shoprite*. Dedecidimos submeter estas variáveis a esta análise exploratória de modo a encontrar as similaridades existente entre elas.

II. *Razões de exclusão de variáveis*

As variáveis que foram excluidas neste estudo são aquelas que já foram tidas como relevantes para serem analizadas nos modelos de *RLB* ou *RLO*, estando desta forma fora do critério de inclusão das variáveis para este estudo.

3 Capítulo II - Tratamento de Dados e Modelos Estatísticos

3.1 Regressão Logística Binária

Segundo Hair et al (2005), a Regressão Logística Binária é um método multivariado, usado para estimar uma variável dependente dicotómica (que assume dois valores), a partir de um conjunto de variáveis independentes, com uma estimação probabilística cujo valor de referência é de 0,5.

Segundo Batistela *et al* (2009), a regressão logística pode ser considerada uma extensão da regressão linear, pois assim como na regressão linear, ela estuda relações entre variáveis, buscando as variáveis que podem influenciar de alguma forma em uma variável dependente, sendo que na regressão logística essa variável dependente deve ser categórica, enquanto a regressão linear dá uma resposta em probabilidade de chances de ocorrer o facto que está sendo estudado.

Segundo Fávero, L. P, & Belfiore, P. (2017, p. 612) a regressão logística binária tem como objectivo principal estudar a probabilidade de um evento definido por Y que se apresenta na forma qualitativa

dicotómica, com base no comportamento das variáveis explicativas. Neste contexto, olhando para as duas grandes redes de supermercados, *Game e Shoprite*, queremos encontrar um modelo de regressão logística binária capaz de prever em termos probabilístico, a probabilidade de um potencial cliente optar por um supermercado de uma ou outra cadeia.

A regressão logística binária tem como objectivo principal estudar a probabilidade de ocorrência de um evento definido por *Y* que se apresenta na forma qualitativa dicotómica *(Y* = 1 para descrever a ocorrência do evento de interesse e *Y* = O para descrever a ocorrência do não evento), com base no comportamento de variáveis explicativas. Desta forma, podemos definir um vector de variáveis explicativas, com respectivos parâmetros estimados, da seguinte forma:

$$Z_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \ldots + \beta_k.X_{ki} \quad (3.1)$$

em que Z é conhecido por **logito, $\boldsymbol{\alpha}$** representa a constante, β_j $(j = 1, 2, \ldots, k)$ são os parâmetros estimados de cada variável explicativa, X são as variáveis explicativas (métricas ou *dummies)* e o subscrito i representa cada observação da amostra ($i = 1, 2, \ldots, n$, em que *n* é o tamanho da amostra). É importante ressaltar que *Z* não representa a variável dependente, denominada por *Y,* e o nosso objectivo neste momento é definir a expressão da **Probabilidade *P;*** de ocorrência do evento de interesse para cada observação, em função do logito *Z,* ou seja, em função dos parâmetros estimados para cada variável explicativa. Para tanto, devemos definir o conceito de **chance**
de ocorrência de um evento, também conhecida por ***odds,*** da seguinte forma:

$$chance\ (odds)_{Y_{i=1}} = \frac{P_i}{1-P_i} \quad (3.2)$$

A regressão logística binária define o logito *Z* como o logaritmo natural da chance, de modo que:

$$ln\left(\ chance_{Y_{i=1}}\right) = Z_i; \quad (3.3)$$

de onde vem que:

$$ln\ \left(\frac{P_i}{1-P_i}\right) = Z_i \quad (3.4)$$

Como o nosso intuito é definir uma expressão para a probabilidade de ocorrência do evento em estudo em função do logito, podemos matematicamente isolar *P;* a partir da expressão (3.4), da seguinte maneira:

$$\frac{P_i}{1-P_i} = e^{Z_i} \tag{3.5}$$

$$P_i = (1 - P_i)e^{Z_i} \tag{3.6}$$

$$P_i(1 + e^{Z_i}) = e^{Z_i} \tag{3.7}$$

Probabilidade de ocorrência do evento:

$$P_i = \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{-Z_i}} \tag{3.8}$$

Probabilidade de ocorrência do não evento:

$$1 - P_i = 1 - \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{Z_i}} \tag{3.9}$$

Obviamente, a soma das expressões (3.8) e (3.9) é igual a 1.

A partir da expressão (3.8), podemos elaborar uma tabela com valores de *p* em função dos valores de *Z.*

A partir das expressões (3.1) e (3.8), podemos definir a expressão geral da probabilidade estimada de ocorrência de um evento que se apresenta na forma dicotómica para uma observação i da seguinte forma:

$$P_i = \frac{1}{1+e^{-(\alpha+\beta_1.X_{1i}+\beta_2.X_{2i}+\ldots+\beta_k.X_{ki})}} \tag{3.10}$$

O que a regressão logística binária estima, portanto, não são os valores previstos da variável dependente, mas sim a probabilidade de ocorrência do evento em estudo para cada observação.

Partiremos, então, para a estimação propriamente dita dos parâmetros do logito.

3.2 Estimação do modelo de regressão logística binária por máxima verosimilhança

Na regressão logística binária, a variável dependente segue uma **distribuição de Bernoulli,** ou seja, o facto de determinada observação *i* ter incidido ou não no evento de interesse pode ser considerado como um ensaio de Bernoulli, em que a probabilidade de ocorrência do evento é *P;* e a probabilidade de

ocorrência do não evento é (1 - *p)*. De maneira geral, podemos escrever que a probabilidade de ocorrência de Y_i, podendo Y_i ser igual a 1 ou igual a 0, é dada por:

$$P(Y_i) = {P_i}^{Y_i}(1 - P_i)^{1-Y_i} \quad (3.11)$$

Para uma amostra com *n* observações, podemos definir a função de verossimilhança *(likelihood function)* como sendo:

$$L = \prod_{i=1}^{n} {P_i}^{Y_i}(1 - P_i)^{1-Y_i} \quad (3.12)$$

de onde vem, com base nas expressões (3.8) e (3.9), que:

$$L = \prod_{i=1}^{n} \left(\frac{e^{Z_i}}{1+e^{Z_i}}\right)^{Y_i} \left(\frac{1}{1+e^{Z_i}}\right)^{1-Y_i} \quad (3.13)$$

Como, na prática, é mais conveniente se trabalhar com o logaritmo da função de verossimilhança, podemos chegar à seguinte função, também conhecida por *log likelihood function:*

$$LL = \sum_{i=1}^{n} \left\{\left[(Y_i).\, ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right)\right] + \left[(1 - Y_i).\, ln\left(\frac{1}{1+e^{Z_i}}\right)\right]\right\} \quad (3.14)$$

E agora cabe uma pergunta: **Quais os valores dos parâmetros do logito que fazem com que o valor de *LL* da expressão (3.14) seja maximizado?** Esta importante questão é a chave central para a elaboração da estimação por máxima verossimilhança (ou *maximum likelihood estimation)* em modelos de regressão logística binária, que se consegue mediante a estimação dos parâmetros $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ com base na seguinte função:

$$LL = max\left\{\sum_{i=1}^{n} \left\{\left[(Y_i).\, ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right)\right] + \left[(1 - Y_i).\, ln\left(\frac{1}{1+e^{Z_i}}\right)\right]\right\}\right\} \quad (3.15)$$

Agora, precisamos verificar se todos os parâmetros estimados são estatisticamente significantes a um determinado nível de significância. Se não for este o caso, precisaremos reestimar o modelo final, a fim de que o mesmo apresente apenas arâmetros estatisticamente significantes para, a partir de então, ser possível a elaboração de inferências e previsões. Portanto, tendo sido elaborada a estimação por máxima verossimilhança dos parâmetros da equação de probabilidade de ocorrência do evento, partiremos para o estudo da significância estatística geral do modelo obtido, bem como das significâncias estatísticas dos próprios parâmetros.

3.3 Significância estatística geral do modelo e dos parâmetros da regressão logística binária

Como a variável dependente é qualitativa, não faz sentido discutirmos o percentual de sua variância que é explicado pelas variáveis preditoras, ou seja, em modelos de regressão logística não há um coeficiente de ajuste R^2 como nos modelos tradicionais de regressão estimados pelo método de mínimos quadrados ordinários. Entretanto, conforme iremos apresentar, existe o critério mais adequado à escolha do melhor modelo, o qual se refere à maior área abaixo da curva ROC.

Muitos pesquisadores também utilizam o pseudo R^2 de McFadden como um indicador de desempenho do modelo escolhido, independentemente da comparação com outros modelos, porém a sua interpretação exige muitos cuidados e, por vezes, há a inevitável tentação em associá-lo, erroneamente, com percentuais de variância da variável dependente. Como iremos apresentar na seção 3.5.3, o melhor indicador de desempenho de um modelo de regressão logística binária refere-se à eficiência global do modelo, que é definida com base na determinação de um *cutoff*, cujos conceitos também serão estudados na mesma seção.

O **teste** X^2 propicia condições à verificação da significância do modelo, uma vez que suas hipóteses nula e alternativa, para um modelo geral de regressão logística, são, respectivamente:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1: existe\ pelo\ menos\ um\ \beta_j \neq 0$$

Enquanto o teste *F* é utilizado para modelos de regressão em que a variável dependente apresenta-se na forma quantitativa, o que gera a decomposição de variância, o teste X^2 é mais adequado para modelos estimados pelo método de máxima verossimilhança, como os modelos de regressão logística. o teste x^2 propicia ao pesquisador uma verificação inicial sobre a existência do modelo que está sendo proposto, uma vez que, se todos os parâmetros estimados β_j (j $= 1, 2, \ldots, k$) forem estatisticamente iguais a zero, o comportamento de alteração de cada uma das variáveis *X* não influenciará em absolutamente nada a probabilidade de ocorrência do evento em estudo. A estatística X^2 tem a seguinte expressão:

$$X^2 = -2(LL_0 - 2.LL_{máx}) \qquad (3.17)$$

Análogo ao teste *F*, o teste X^2 avalia a significância conjunta das variáveis explicativas, não definindo qual ou quais destas variáveis consideradas no modelo são estatisticamente significantes para influenciar a probabilidade de ocorrência do evento.

Desta forma, é preciso que o pesquisador avalie se cada um dos parâmetros do modelo de regressão logística binária é estatisticamente significante e, neste sentido, a **estatística *z* de Wald** será importante para fornecer a significância estatística de cada parâmetro a ser considerado no modelo. As hipóteses do **teste *z* de Wald** para o α e para cada β_j são, respectivamente:

$H_0: \alpha = 0$

$H_1: \alpha \neq 0$

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

As expressões para o cálculo das estatísticas ***z* de Wald** de cada parâmetro α e β_j. são dadas, respectivamente, por:

$$z_\alpha = \frac{\alpha}{s.e.(\alpha)}$$

$$z_{\beta_j} = \frac{\beta_j}{s.e.(\beta_j)} \qquad (3.18)$$

em que *s.e.* significa o erro-padrão *(standard error)* de cada parâmetro em análise.

Para tanto, o **teste de razão de verossimilhança *(likelihood-ratio test),*** que verifica a adequação do ajuste do modelo completo em comparação com o ajuste do modelo final, pode ser utilizado, apresentando a seguinte expressão:

$$X^2{}_{1\ g.l.} = -2\left(LL_{modelo\ final} - 2.LL_{modelo\ completo}\right) \qquad (3.19)$$

Conforme podemos perceber, estes cálculos utilizaram sempre as estimativas médias dos parâmetros.

3.4 *Cutoff,* análise de sensibilidade, eficiência global do modelo, sensitividade e especificidade

Estimado o modelo de probabilidade de ocorrência do evento, vamos agora definir o conceito de *cutoff*, a partir do qual será possível classificar, as observações com base nas probabilidades estimadas de cada uma delas.

O *cutoff*, que nada mais é do que um ponto de corte que o pesquisador escolhe, é definido para que sejam classificadas as observações em função das suas probabilidades calculadas e, desta forma, é utilizado quando há o intuito de se elaborarem previsões de ocorrência do evento para observações não presentes na amostra com base nas probabilidades das observações presentes na amostra.

Assim, se determinada observação não presente na amostra apresentar uma probabilidade de incidir no evento maior do que o *cutoff* definido, espera-se que haja a incidência do evento e, portanto, será classificada como *evento.* Por outro lado, se a sua probabilidade for menor do que o *cutoff* definido, espera-se que haja a incidência do não evento e, portanto, será classificada como *não evento.*

De maneira geral, podemos estipular o seguinte critério:

Se $P_i > cutoff$ a observação i deverá ser classificada como *evento.*

Se Se $P_i < cutoff$ a observação i deverá ser classificada como *não evento.*

Como a expressão de probabilidade é estimada com base nas observações presentes na amostra, a classificação para outras observações não presentes inicialmente na amostra leva em consideração a consistência do comportamento dos estimadores e, portanto, para efeitos inferenciais, a amostra deve ser significativa e representativa do comportamento populacional, como em qualquer modelo de dependência confirmatório.

O *cutoff* serve para que o pesquisador avalie a real incidência do evento para cada observação e a compare com a expectativa de que cada observação incida, de fato, no evento. Com isto feito, será possível avaliar a taxa de acerto do modelo com base nas próprias observações presentes na amostra e, por inferência, assumir que tal taxa de acerto se mantenha quando houver o intuito de avaliar a incidência do evento para outras observações não presentes na amostra (previsão).

Esta análise, conhecida por **análise de sensibilidade**, gera classificações que dependem da escolha do *cutoff.*

Neste momento, definiremos os conceitos de **eficiência global do modelo**, **sensitividade e especificidade**.

A **eficiência global** do modelo corresponde ao percentual de acerto da classificação para um determinado *cuteff.* A eficiência global do modelo é calculada da seguinte forma:

$$EGM = \frac{número\ \ previsões\ correcta}{úmero\ total\ de\ previsões} \tag{3.21}$$

Conforme mencionado na seção 3.5.2, a eficiência global do modelo, para um determinado *cutoff*, é bem mais adequada para se avaliar o desempenho da modelagem do que o pseudo R^2 de McFadden, uma vez que a variável dependente apresenta-se na forma qualitativa dicotômica.

A **sensitividade** diz respeito ao percentual de acerto, para um determinado *cutoff*, considerando-se apenas as observações que de facto são evento.

$$Sensitividade = \frac{verdadeiros\ positivos}{verdadeiros\ positivos + falsos\ negativos} \quad (3.22)$$

Já a **especificidade**, por outro lado, refere-se ao percentual de acerto, para um dado *cutoff*, considerando-se apenas as observações que não são evento. A sua expressão é dada por:

$$Especificidade = \frac{verdadeiros\ positivos + verdadeiros\ negativos}{total\ de\ observações} \quad (3.23)$$

Em trabalhos acadêmicos e em relatórios gerenciais de diversas organizações, é comum que sejam apresentados e discutidos alguns gráficos da análise de sensibilidade. Os mais comuns são os conhecidos por **curva de sensibilidade e curva *ROC*** *(Receiver Operating Characteristic),* que apresentam finalidades distintas. Enquanto a curva de sensibilidade é um gráfico que apresenta os valores da sensitividade e da especificidade em função dos diversos valores de *cutoff,* a curva *ROC* é um gráfico que apresenta a variação da sensitividade em função de (1 - especificidade).

Por meio da curva de sensibilidade, podemos verificar que é possível definir o *cutoff* que iguala a sensitividade com a especificidade, ou seja, o *cutoff* que faz com que a taxa de acerto de previsão para aqueles que serão evento seja igual à taxa de acerto para aqueles que não serão evento. É importante mencionar, contudo, que este *cutoff* não garante que a eficiência global do modelo seja a máxima possível.

Além disso, a curva de sensibilidade perrmite que o pesquisador avalie o *trade off* entre sensitividade e especificidade quando da alteração do *cutoff*,já que, em muitos casos, conforme discutido,o objetivo da previsão pode ser o de aumentar a taxa de acerto para aqueles que serão evento sem que haja uma perda considerável de taxa de acerto para aqueles que não serão evento.

A curva *ROC* mostra o comportamento propriamente dito do *trade of* entre sensitividade e especificidade e, ao trazer, no eixo das abscissas, os valores de (1 - especificidade), apresenta formato convexo em relação ao ponto (0, 1). Desta forma, um determinado modelo com maior área abaixo da curva *ROC* apresenta maior eficiência global de previsão, combinadas todas as possibilidades de *cutoff*

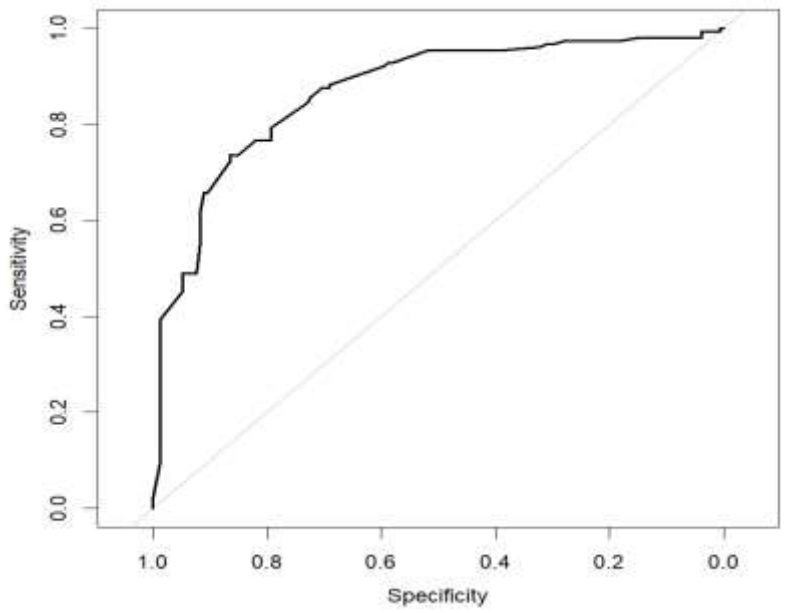
e, assim, a sua escolha deve ser preferível quando da comparação com outro modelo com menor área abaixo da curva *ROC.* Em outras palavras, se um pesquisador desejar, por exemplo, incluir novas variáveis explicativas na modelagem, a comparação do desempenho global dos modelos poderá ser elaborada com base na área abaixo da curva ROC,já que, quanto maior a sua convexidade em relação ao ponto (0, 1), maior a sua área (maior sensitividade e maior especificidade) e, consequentemente, melhor o modelo estimado para efeitos de previsão.

Segundo Swets (1996), a curva *ROC (Receiver Operating Characteristic),* possui este nome porque compara o comportamento de alteração de duas características operacionais do modelo (sensitividade e especificidade).

Foi primeiramente desenvolvida e utilizada por engenheiros na Segunda Guerra Mundial quando do estudo para detecção de objetos inimigos em batalhas. Na sequência, foi logo introduzida na Psicologia para a investigação das detecções perceptuais de determinados estímulos e, actualmente, é bastante utilizada em campos da Medicina, como a radiologia, e em diversos campos das ciências sociais aplicadas, como Economia e Finanças.

Neste caso específico, é consideravelmente utilizada em modelos de gestão de risco de crédito e de probabilidade de *default.*

Ilustração 1: Curva ROC



Fonte:Autor

3.5 Índice Kappa de Cohen

O Teste de Kappa (sugerido por Cohen em 1960) é uma medida de concordância interobservador e mede o grau de concordância além do que seria esperado tão somente pelo acaso.

Para descrevermos se há ou não concordância entre dois ou mais avaliadores, ou entre dois métodos de classificação, utilizamos a medida Kappa que é baseada no número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre os avaliadores. No caso da nossa pesquisa queremos avaliar a concordância entre os resultados ou previsões do modelo e as observações. Esta medida de concordância assume valor máximo igual a 1, que representa total concordância ou ainda pode assumir valores próximos e até abaixo de 0, os quais indicam nenhuma concordância.

O coeficiente Kappa é calculado a partir da seguinte fórmula:

$$K = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \qquad (3.24)$$

$\sum \pi_{ii}$ – probabilidade de concordância

$\sum \pi_{i+}\pi_{+i}$ – Concordância esperada se as observações fossem esperadas

π_{i+} - proporção de ocorrência da categoria i para o avaliador (modelo)

π_{+i} - proporção de ocorrência da categoria i para o avaliador(Observações)

Interpretação dos índices de Kappa

Valores obtidos de Kappa	Interpretação
<0	Nenhuma concordância
0-0,19	Concordância pobre
0,20-0,39	Concordância leve
0,40-0,59	Concordância moderada
0,60-0,79	Concordância substancial
0,80-1,00	Concordância quase perfeita a perfeita

3.6 Regressão Logística Ordinal

Segundo Agresti (2012), muitas situações de regressão categórica quando a variável dependente assume mais de duas categorias mutuamente exclusivas ordinais, a regressão multinomial não tem em conta essa

relação de ordem entre elas. A modelação que consiste na utilização do carácter ordinal da variável dependente para fazer inferências, que colmata a insuficiência da regressão multinomial, chama-se regressão logística ordinal ou, simplesmente, regressão ordinal.

Exemplos de variáveis categóricas ordinais encontram-se em pesquisas em epidemiologia, onde se deseja predizer o estágio ou nível de gravidade de uma doença (leve, moderada, grave), pesquisas em educação, onde procura aferir-se o grau de proficiência numa língua (elementar, independente, proficiente), e pesquisas em educação, onde se deseja predizer o grau de soluções de problemas aritméticos (fraco, aceitável, excelente).

A regressão ordinal, à semelhança do modelo de regressão multinomial, é um modelo de probabilidades, mas agora a ocorrência de uma categoria é expressa em termos de probabilidades acumuladas ou cumulativas das várias categóricas da variável dependente. Assim, este modelo é também chamado de modelo de probabilidades cumulativas.

3.7 Modelos de Regressão Logística Ordinal

Um modelo de regressão logística ordinal pode ser desenvolvido de várias formas distintas. A abordagem a usar para um modelo de regressão logística ordinal (MRLO) é muito semelhante à do modelo de regressão logística binária (MRLB). De facto o modelo MRLB pode ser visto como um caso especial do modelo ordinal para o qual a variável de resposta somente possui duas categorias.

O M RLO pode ser expressa como um modelo de variável latente (Agresti, 2002; Long e Freese, 2001). Considerando a existência de uma variável latente, Y^*, pode-se definir $Y^* = \vec{x}\beta + \varepsilon$, sendo $\vec{x}$ um vector linha $(1 * k)$ e β um vector coluna $(k * 1)$ de coeficientes estruturais, e ε corresponde a uma perturbação aleatória com uma distribuição normal reduzida, i.e. $\varepsilon \sim N(0, 1)$.

Assumindo que a variável latente Y^* é definida como função do conjunto de variáveis explicativas e do erro aleatório, pode-se considerar que esta variável pode tomar um conjunto infinito de valores, os quais podem ser colapsados num conjunto de categorias da variável de resposta Y. A variável latente Y^* vai ter vários pontos de corte (limites): α_1, α_2, ...,α_j e o valor da variável observada y estará dentro das regiões definidas por esses pontos de corte. Considere-se, a título de exemplo, a variável ordinal dependente "Sente-se satisfeito com este Supermercado", consistindo num conjunto de cinco categorias ou cinco níveis e variando de "pouco satisfeito" a "muito satisfeito". Sendo o nível de satisfação uma variável ordinal, y, variando de 1 a 5, onde 1 ="pouco satisfeito", 2="pouco", 3= "mais ou menos", 4="muito" e 5="muito satisfeito" , define-se os pontos de corte de tal forma que $\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$.

$$Y = \begin{cases} 1 \; se \; Y^* < \alpha_1 \\ 2 \; se \; \alpha_1 < Y^* \leq \alpha_2 \\ 3 \; se \; \alpha_2 < Y^* \leq \alpha_3 \\ 4 \; se \; \alpha_3 < Y^* \leq \alpha_4 \\ 5 \; se \; \alpha_4 < Y^* \leq +\infty \end{cases} \tag{3.25}$$

Pode-se calcular a probabilidade para cada nível de satisfação. Por exemplo:

$$\begin{aligned} \pi_1 &= P(y = 1) = P(y^* < \alpha_1) = P(x\beta + \varepsilon \leq \alpha_1) = F(\alpha_1 - x\beta) \\ \pi_2 &= P(y = 2) = P(\alpha_1 < y^* \leq \alpha_2) = F(\alpha_2 - x\beta) - F(\alpha_1 - x\beta) \\ \pi_3 &= P(y = 3) = P(\alpha_2 < y^* \leq \alpha_3) = F(\alpha_3 - x\beta) - F(\alpha_2 - x\beta) \\ \pi_4 &= P(y = 4) = P(\alpha_3 < y^* \leq \alpha_4) = F(\alpha_4 - x\beta) - F(\alpha_3 - x\beta) \\ \pi_5 &= P(y = 5) = P(\alpha_4 < y^* \leq +\infty) = 1 - F(\alpha_4 - x\beta) \end{aligned} \tag{3.26}$$

Seja Y a variável de resposta com k categorias codificadas de 1, 2, 3, …, k, e o vetor de variáveis explicativas definido por $\vec{x} = (x_1 \, , \, x_2, \dots, x_p)$. As k categorias da variável de resposta Y tendo em conta as covariáveis consideradas ocorrem com probabilidades π_1, π_2,..., π_k ou seja $\pi_j = P(y = j)$, para j = 1, 2, …, k.

Também se pode calcular as probabilidades cumulativas usando a fórmula:

$$P(Y \leq j) = F(\alpha_j - x\beta) \qquad \text{com j = 1, 2, ..., j-1} \tag{3.27}$$

3.8 Modelo de odds proporcionais (OP)

O modelo de logit cumulativo foi originalmente proposto por Walker e Duncan (1967) e mais tarde chamado de modelo de odds proporcionais por McCullagh (1980). A dependência de Y sobre $\vec{x} = (x_1 \, , \, x_2, \dots, x_p)$, para o modelo de odds proporcionais, pode ser representado da seguinte forma:

$$[\pi(Y \leq j | x_1 \, , x_2, \dots, x_p)] = \frac{e^{\alpha_j + (-\beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p)}}{1 + e^{\alpha_j + (-\beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p)}} = \frac{e^{(\alpha_j - \beta x)}}{1 + e^{(\alpha_j - \beta x)}} \qquad \text{com j=1, 2, ..., k} \tag{3.28}$$

No modelo de odds proporcionais consideram-se (k – 1) pontos de corte das categorias, sendo que o j-ésimo ponto de corte é baseado na comparação de probabilidades acumuladas. No modelo de odds

proporcionais trabalha-se com o logit, ou seja, com o logaritmo natural dos odds. Para estimar o ln(odds) de estar numa dada categoria ou abaixo dela o modelo de odds proporcionais pode ser escrito na seguinte forma:

$$logit\pi(\vec{x}) = ln\left[\frac{\pi_j(\vec{x})}{1-\pi_j(\vec{x})}\right] = logit\left[\pi\left(Y \leq j|x_1, x_2, \dots, x_p\right)\right] = ln\left[\frac{\pi\left(Y \leq j|x_1, x_2, \dots, x_p\right)}{\pi\left(Y > j|x_1, x_2, \dots, x_p\right)}\right] =$$

$$= \alpha_j + \left(-\beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p\right) \quad (3.29)$$

Onde $\pi_j(\vec{x}) = \pi\left(Y \leq j|x_1, x_2, \dots, x_p\right)$ representa a probabilidade de estar na categoria j ou abaixo dela, distribuição cumulativa de probabilidades, dado o conjunto de variáveis explicativas considerado.

Pode parecer confuso o facto de o modelo subtrair βx em vez de adicionar. Isto resulta do facto de se calcular a probabilidade de y ≤ j em vez de y > j.

Os α_j são os parâmetros desconhecidos de intersecção que satisfazem a condição $\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \dots \leq \alpha_k$ é o vetor dos coeficientes de regressão desconhecidos correspondentes $\vec{x} = \left(x_1, x_2, \dots, x_p\right)$.

Por transformação dos *logit* cumulativos podemos obter os *odds* cumulativos assim como as probabilidades cumulativas de estar na categoria j ou abaixo dela. Com base no ajuste do modelo a razão de odds cumulativos Ψ_l para a covariável binária de ordem *l*, representada por x_l pode ser obtida da seguinte forma:

$$\Psi_{op} = \frac{\pi\left(Y \leq y_j|X_l^{(1)}\right)}{\pi\left(Y \leq y_j|X_l^{(0)}\right)} = e^{\left\{-\beta_l\left(X_l^{(1)} - X_l^{(0)}\right)\right\}} \quad (3.30)$$

O modelo assenta no pressuposto de *odds* proporcionais acerca dos (k – 1) pontos de corte.

O pressuposto de *odds* proporcionais também é conhecido por pressuposto de regressão paralela que é assumida para cada covariável incluída no modelo. De notar que o vetor dos coeficientes de regressão, β, não depende de *j* o que implica que o modelo assume que a relação entre x_j e Y é independente de j. McCullagh (1980) denomina este pressuposto de odds proporcionais para os (k – 1) pontos de corte, também chamado de pressuposto de regressão paralela, que é assumida para cada covariável incluída no modelo, e daí o nome do modelo, pois assume-se que a razão de *odds* de qualquer variável explicativa é constante ao longo de todas as categorias ordenadas.

3.9 Índice de Kappa ponderado para dois avaliadores

Dada a necessidade de diferenciar o grau de discordância entre as diferentes categorias quando se trabalha com variáveis com escala ordinal, surgiu a ideia de atribuir pesos diferentes a essas categorias

discordantes. Quanto mais afastada estiver a categoria discordante da categoria concordante, menor será o peso atribuído a essa categoria.

Por exemplo, supondo que tenhamos cinco categorias para medir a satisfação de um cliente relativamente aos serviços fornecidos por uma supermercado: “insatisfeito”, “ pouco satisfeito”,” indiferente”,” satisfeito” e “muito satisfeito” e dois avaliadores (podendo ser uma previsão e outra observação). Neste caso, a discordância entre um avaliador classificar como "muito satisfeito" enquanto outro classificar como “satifeito" não será muito relevante, mas se um classificar como "muito satisfeito” enquanto o outro categorizar como “insatisfeito” (ou seja, nos opostos da escala), esta discordância será mais relevante. Nestas situações o Kappa proposto por Cohen (1968) é ineficiente para analisar avaliações medidas numa escala ordinal. O próprio autor propôs a versão ponderada do kappa para corrigir esse problema, (Cohen, 1968). A passagem de um coeficiente kappa (K) para um coeficiente Kappa ponderado (KW) permite atribuir diferentes pesos às discordâncias, tornando-se assim uma estatística preferível para dados com categorias ordenadas.(Cohen, 1968).

O kappa ponderado atribui menos peso para o acordo quando as categorias estão mais afastadas. Uma discordância de “muito satisfeito” versus “indiferente” ainda seria onsiderado um acordo parcial, mas um desacordo de “muito satisfeito” versus "insatisfeito" seria contado como um total desacordo, sendo atribuído um peso muito baixo.

O Kappa ponderado é então um índice estatístico utilizado para determinar a fiabilidade quando as variáveis são ordinais e os resultados podem ser expressos por mais de duas categorias sendo considerado uma extensão do kappa de Cohen (dado que este pode ser utilizado em situações com variáveis nominais/categóricas ou ordinais). Enquanto o Kappa de Cohen não ponderado considera somente concordância ou discordância, o Kappa ponderado permite a atribuição de pesos às diferentes categorias, de tal forma que categorias semelhantes podem estar em acordo parcial.(Cohen, 1968)

O kappa ponderado tem as mesmas limitações dos Kappas não ponderados. Esta estatística é adequada quando temos entre 3 e 10 categorias ordinais e o tamanho mínimo da amostra necessário para se poder aproximar a uma distribuição normal é de $2 * q^2$ onde q é o número de categorias (Domenic V. Cicchetti & Feinstein, 1990). Os pesos estão compreendidos num intervalo $0 \leq w_{kl} \leq 1$, onde $k = 1,2, \dots, q$ e $l = 1,2, \dots, q$. O peso máximo será atribuído quando o acordo entre os dois avaliadores é exato, isto é, $w_{kk} = 1$, e a todos os desacordos será atribuído um peso com um valor inferior ao peso máximo(D. V. Cicchetti, 1981).

Para o nosso trabalho, dada a característica das nossas variáveis que são ordinais, (Gwet, 2010) sugere a utilização de pesos ordinais definidos através da relação:

$$w_{kl} = \begin{cases} 1 - \frac{M_{kl}}{M_{max}}, & se\ k \neq l \\ 1 \qquad\quad , & se\ \ k = l \end{cases} \tag{3.31}$$

Para o cálculo do Kappa de Cohen ponderado (weighted Kappa, K_{CW}) é necessário calcular a proporção de concordância observada ponderada (*Paw*) e a proporção de acordo devido ao acaso ponderado (P_{ew}). O cálculo (P_{aw}) é dado pela seguinte relação:

$$p_{aw} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} p_{kl} \tag{3.32}$$

onde as proporções p_{kl} representam as avaliações (concordantes e discordantes) dadas entre os dois avaliadores. O valor obtido desta relação é interpretado como a percentagem de concordância ponderada.

A proporção ponderada de acordo devida ao acaso (*Pew*) é dada por:

$$p_{ew} = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} p_{k+} p_{+l} \tag{3.33}$$

onde as proporções p_{k+} e p_{+l} representam as respetivas frequências marginais. Consequentemente, o Kappa de Cohen ponderado é dado por:

$$k_{cw} = \frac{p_{aw} - P_{ew}}{1 - P_{ew}} \tag{3.34}$$

Convém salientar que quando todos os desacordos são considerados igualmente graves, ou seja w_{kl}=0 para todo o $k \neq l$ e $w_{kl} = 1$ para todo o $k = l$, então o kappa ponderado é idêntico ao kappa não ponderado tratado no parágrafo 3.4.

3.10 Escalonamento Multidimensional – MDS

3.10.1 Medidas de Proximidade

Introdução

Quando uma análise de dados requer a aplicação de alguma técnica de Escalonamento multidesional – MDS (do inglês *Multidimensional Scaling*), a primeira etapa a ser considerada é a definição da medida de proximidade a ser usada. Na prática os dados podem apresentar-se tanto sob a forma de uma matriz

de dados que apresenta observações de *p* objetos feitas por *n* individuos, como sob a forma de uma matriz de proximidade, que apresenta uma medida de proximidade para cada par de objetos.

Quando os dados são coletados na forma de medidas de proximidade a aplicação da técnica de MDS é imediata, mas quando os dados são coletados sob aforma de uma matriz de dados (*n x p*) é necessário que se obtenha medidas de proximidade derivadas desses dados, afim de que se possa aplicar uma técnica de MDS.

Uma medida de proximidade, referida também como medida de similaridade , dissimilaridade ou distância, deve ser um valor que indique o quanto dois objetos são semelhantes (medida de similaridade), ou o quanto dois objetos são diferemtes ou estão distantes um do outro (medida de dissimilaridade ou distância), ou ainda um valor que indique o grau com que dois objetos são percebidos como sendo semelhantes ou diferentes por determinado individuo (por exemplo, uma matriz de correlação).

Uma das formas mais comuns de se derivar uma medida de proximidade é calculando coeficientes de similaridade ou dissimilaridade entre os objectos. Uma vez obtidas as medidas de proximidade elas são dispostas em matrizes nas quais cada elemento δ_{ij} representa a proximidade entre objecto i e objecto j.

Medidas de similaridade e dissimilaridade são rigorosamente relacionadas de uma maneira inversa, ou seja se δ é uma função definida sobre cada par de objectos, a qual mede a similaridade entre objectos, então é fácil derivar uma medida correspondente de dissimilaridade assim com $\delta^* = (contante - \delta)$.

Medidas de Proximidade Para Dados Quantitativos

Considere uma matriz de dados **(n** x p) apresentando **n** observações de p objectos .

A mais familiar medida de dissimilaridade entre objectos é a distância Euclidiana, tal que

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \qquad (3.35$$

A distância *Euclidiana* usada em dados brutos pode ser muito insatisfatória quando as variáveis são medidas em diferentes unidades e tem variâncias diferentes e também se as variáveis são correlacionadas , visto que a distância *Euclidiana* é muito afectada por mudança de escala, podendo mudar não só os valores das didtâncias assim como o posto, a distância é usualmente calculada após a padronização das variáveis.

Existem outras métricas que têm sido usadas e uma delas é a métrica (distancia) de *Minkowski*, que é dada pela fórmula

$$d_{ij} = \sqrt[R]{\sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right|^{R}} \qquad (R \geq 0) \tag{3.36}$$

Sendo a métrica *Euclidiana* um caso particular de *Minkowski* para $n = 2$.

Para $R = 1$ temos a métrica de *Manhatan*, tambem conhecida por métrica absoluta ou "city block", dada pela fórmula:

$$M_{ij} = \sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right| \tag{3.37}$$

A mais usual medida de similaridade entre duas variáveis (ou dois objectos) é o coeficiente de correlação de Pearson, tambem conhecido como correlação momento-producto que é dada por

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \tag{3.38}$$

Uma outra medida de similaridade que pode ser usada é o cosseno do ângulo entre vectores dado por

$$\cos\theta = \frac{x_i^T x_j}{\left[x_i^T x_i\right]\left[x_j^T x_j\right]} \tag{3.39}$$

Existem outras medidas de similaridades, dissimilaridades e distâncias, que podem ser usadas mas estas são as mais conhecidas e comumente usadas.

Medidas de Proximidades Para Variáveis Qualitativas

O coeficiente de correlação de Kendall

O coeficiente de correlação de Kendall, também conhecido como tau de Kendall, é uma medida não paramétrica da associação entre duas variáveis ordinais. Ele mede a concordância entre duas classificações, levando em conta o número de pares concordantes e discordantes

3.10.2 Fórmula do Coeficiente de Correlação de Kendall (τ))

A fórmula geral para o cálculo de coeficiente de correlação de Kendall é dada por:

$$\tau = \frac{C-D}{\sqrt{(C+D+T)\cdot(C+D+U)}} \quad (3.40)$$

Onde:

- C é o número de pares concordantes.
- D é o número de pares discordantes.
- T é o número de pares empatados em uma das variáveis.
- U é o número de pares empatados na outra variável.

Fórmula Simples do Coeficiente de Kendall (τ)

Se não houver empates, a fórmula simplificada é:

$$\tau = \frac{C-D}{\frac{1}{2}n(n-1)} \quad (3.41)$$

Onde:

- C é o número de pares concordantes.
- D é o número de pares discordantes.
- n é o número de observações.

Definição de Pares Concordantes e Discordantes

- **Par Concordante**: Para dois pares de observações (x_i, y_i) e (x_j, y_j), eles são concordantes se a ordem relativa entre x_i e x_j é a mesma que a ordem relativa entre y_i e y_j . Isto é, se $x_i > x_j$, então $y_i > y_j$ ou $x_i < x_j$, então $y_i < y_j$.
- **Par Discordante**: Para dois pares de observações (x_i, y_i) e (x_j, y_j), eles são discordantes se a ordem relativa entre x_i e x_j é oposta da ordem relativa entre y_i e y_j . Isto é, se $x_i > x_j$, então $y_i < y_j$ ou $x_i < x_j$, então $y_i > y_j$.

3.10.3 Métodos de Escalnamento Multidimensional

Introdução

Os dois tipos mais importantes de procedimentos de escalonamento são denominados escalonamento métrico (também referido como escalonamento clássico ou Análise de Coordenadas Principais) e escalonamento não métrico (também referido como escalonamento ordinal multidimensional), ou, simplesmente, escalonamento multidimensional.

Modelos clássicos de escalonamento multidimnesional foram designados para preservar informação métrica dos dados, assumindo que as proximidades eram uma função linear do modelo de distâncias, enquanto modelos de escalonamento multidimensional não métricos procuravam somente preservar a informção métrica dos dados. Em procedimentos de escalonamento multidimensional esta distinção tem sido implementada, basicamente, pela forma de regressão usada, sendo que, regressão linear de dados sobre distâncias tem sido usada no caso métrico e regressão monótona tem sido usada no caso não métrico.

O escalonamento não métrico usa as verdadeiras magnitudes das proximidades originais entre os objectos para obter uma representação geométrica destes objectos, e sendo assim, deve ser usado quando os objectos a serem escalonados estiverem em escala de razão ou na escala intervalar.

Por outro lado, o escalonamento não métrico usa somente as propriedades ordinais das proximidades originais entre objectos e, assim sendo, deve ser usado quando os dados estiverem na escala ordinal.

Método de Escalonamento Multidimensional Métrico

O escalonamento métrico é um método de representação algébrico que tem como objectivo achar uma configuração de pontos a partir das dissimilaridades entre os objectos, o qual é apropriado particularmente quando as dissimilaridades são, exactamente ou aproximadamente, distâncias euclidianas.

Este método admite a existência de uma configuração de *k* dimensões, cujas distâncias entre os pontos são δ_{ij} e tenta reconstruir esta configuração usando uma matriz de distâncias observadas D, cujos elementos são da forma:

$$d_{ij} = \delta_{ij} + e_{ij}, \qquad (5.8)$$

onde e_{ij} são erros de medida acrescidos de erros de distorções causados pelo facto das distâncias observadas não corresponderem exactamente às distâncias entre os pontos de uma configuração R^k.

Deve-se notar que, dada uma série de distâncias euclidianas, não existe uma única representação dos pontos os quais dão origem a estas distâncias, isto significa que não se pode determinar a localização e a orientação da configuração. O problema de localização é usualmente superado centrando a configuração na origem, enquanto que o problema da orientação pode ser superado submetendo a configuração obtida a uma transformação ortogonal, a qual deixa as distancias e ângulos inalterados. Em outras palavras, uma configuração obtida pelo escalonamento métrico é indeterminada com respeito a translações, rotações e reflexões.

Método de Escalonamento Multidimensional não Métrico (Sheppard e Kruskall)

O escalonamento não-métrico ou ordinal assim como o escalonamento métrico, têm como propósito encontrar uma representação geométrica de pontos cujas distâncias entre estes pontos combinem de alguma maneira com as similaridades (ou dissimilaridades) originais entre objectos (ou individuos) .

A diferença básica entre os métodos é que o não-métrico usa somente a ordenação das similaridades para obter a solução, enquanto que o métrico usa informação métrica, ou seja, as verdadeiras magnitudes das similaridades, para obter a solução. O escalonamento não-métrico, entre tanto, toma como hipótese uma relação menos rigida, do que a assumida pelo métrico, entre d_{ij} e δ_{ij},

$$d_{ij} = f(\delta_{ij}) + e_{ij} \qquad (5.9)$$

sendo f uma função monótona desconhecida crescente ou decrescente. Em ouras palavras, o escalonamento não-métrico relaxa a forte linearidade suposta pelo escalonamento métrico a respeito do tipo de função que liga as similaridades entre objectos às distâncias entre os pontos, que representam os objectos na configuração.

A grande vantagem do uso de escalonamento não-métrico é que o mesmo algoritmo básico é facilmente generalizado para diferentes tipos de dados e para diferentes modelos. É, portanto, aplicável a uma grande variedade de situações. Assim, frequências, Probabilidades, postos, etc, são tão apropriados como medidas de proximidade, como são os ***coeficientes de correlação***, associação, covariâncias e outros. Obviamente, quaisquer séries de medidas com a mesma ordenação de proximidades irão gerar a mesma solução métrica.

Como em escalonamento métrico, não se pode determinar uma única localização e uma única orientação da configuração obtida, isto é, a configuração é indeterminada com respeito a translação, rotação e reflexão. Além disso, a configuração obtida pelo método não-métrico de escalonamento é indeterminada

com respeito a mudanças de escala, isto é, é indeterminada com respeito a expansão ou contração uniformes, entretanto, o problema de não ter signicado a escala da configuração, pode ser
superado fixando o centróide da configuração na origem, isto é, requerendo que a raiz quadrada da média das distâncias dos pontos da origem seja unitária.

Para o nosso trabalho usaremos uma das opções sugeridas para o escalonamento não métrico, que é o uso da matriz das correlações, para o cálculo das similaridades. Neste caso, e por se tratar de MDS para variáveis ordinais, usaremos os coeficientes de correlação de *Kendall*,como sugere Khamis, H. no seu artigo *"measures of association how to choose "*, 2008.

Adequação de Ajuste

Com o método de escalonamento não-métrico, proposto por Kruskal, é centrado na medida de adequação de ajuste, denominada ***Stress – Standardized Residual Sum of Squares***, considerou-se imporpante que essa medida fosse inicialmente apresentada.

O escalonamento não-métrico requer que as proximidades observadas entre objectos concordem, de alguma maneira, com as distâncias entre pontos da configuração obtida, ou seja, requer uma relação entre proximidades e distâncias do tipo

$$d_{ij} = f(\delta_{ij}) + e_{ij}$$

sendo *f* uma função monótona qualquer .

Surge então a necessidade de definir uma função, a qual possa fornecer uma medida da bondade de ajuste (ou maldade de ajuste) das distâncias do espaço de configuração para com as proximidades observadas e, consequentemente, uma medida do ajuste da configuração obtida aos dados.

Esta função deve assumir o valor zero quando o modelo das distâncias do espaço de configuração ajustar-se perfeitamente ao modelo das proximidades e deve assumir valores maiores quando o ajuste vier a ser menos perfeito.

A função proposta, inicialmente, para medir o ajuste de qualquer configuração, foi a soma de quadrados de desvios, definida por:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij} - f(\delta_{ij})\right)^2 \qquad (3.42)$$

Uma vez que, a monotonia nem sempre é obtida na configuração, define-se os números

$\hat{d}_{ij} = \hat{f}(\delta_{ij})$, os quais são monotonicamente relacionados com as proximidades observadas.

Então, tem-se,

$$d_{ij} = \hat{d}_{ij} + e_{ij} \tag{3.43}$$

e a medida do ajuste vem a ser

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij} - \hat{d}_{ij}\right)^2 \tag{3.44}$$

Embora esta medida seja invariante com respeito a translações, reflexões e rotaçõe, não é invariante com respeito a extensões e contrações uniformes.

Introduziu-se, então, um fator de escala, o qual não só normaliza a medida como também a torna invariante com respeito a mudanças de escala. Este factor de escala é definido por

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij}\right)^2 \tag{3.45}$$

A função que mede o ajuste de qualquer configuração pode ser obtida, então, por

$$\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij}\right)^2} \tag{3.46}$$

Kruskal propôs uma medida de adequação de ajuste definida por

$$\left[\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(d_{ij}\right)^2}\right]^{\frac{1}{2}} \tag{3.47}$$

a qual ele chamou de "stress".

O "stress" é considerado uma função das (*nxk*) coordenadas dos pontos da configuração obtida e, sendo assim, o valor do *stress* depende da configuração obtida. Sabe-se que o valor do *stress* aumenta quando *n* aumenta e/ou quando *k* diminui.

Pode-se dizer, ainda, que o *stress* representa a extensão para a qual os postos de ordenação dos d_{ij} não combinam com os postos de ordenação dos postos de δ_{ij} e, sendo assim, quando os postos dos d_{ij} combinam com os postos de ordenação dos postos de δ_{ij}, o vlor de *stress* é zero.

Kruskal sugere que o valor do *stress* obtido , seja informalmente interpretado de acordo com o seguinte padrão:

Tabela 1: Escala classificativa da qualidade de stress (regra de ouro)

Stress	Qualidade
$> 0{,}200$	Insuficiente
$]0{,}100;\ 0{,}200]$	Regular
$]0{,}050;\ 0{,}00]$	Bom
$]0{,}025;\ 0{,}050]$	Excelente
$\leq 0{,}025$	Perfeito

Fonte: Kruskal, J. B. 1964 *(Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis)*

Decomposição de Stress

A decomposição de stress é uma medida de avaliação que quantifica o grau de distorção entre as distâncias originais (medidas de similaridade ou dissimilaridade) e as distâncias representadas no espaço de baixa dimensão pelo MDS. Quanto menor o valor do stress, melhor é a representação dos dados. Assim, a decomposição de stress é uma ferramenta útil para avaliar o quão bem o MDS está representando a estrutura dos dados originais em um espaço de menor dimensão, permitindo ao o ajuste correcto e optimização dos parâmetros do modelo para obter uma representação mais fiel e útil dos dados.

A decomposição de stress é feita para analisar e entender as fontes de erro na representação de baixa dimensão gerada pelo MDS. Ela pode ajudar a identificar quais componentes contribuem mais para o erro total de representação.

Aqui estão as principais fórmulas envolvidas:

a) *Erro Quadrático Médio (EQM):*

Esta é a medida geral da variação nos dados que não pode ser explicada pelo modelo.

$$EQM = \frac{\sum_{i=1}^{n}(d_{ij} - \hat{d}_{ij})^2}{n} \qquad (3.48)$$

Onde, d_{ij} são as distâncias reais entre os pontos no espaço multidimensional e $\hat{d}_{ij}$ são as distâncias estimadas pelo modelo.

b) *Contribuição do Stress (Stress):*

O Stress é uma medida que indica o quão bem o modelo de MDS representa as distâncias originais.

$$Stress = \sqrt{\frac{EQM}{\sum_{i=1}^{n} {d_{ij}}^2}} \tag{3.49}$$

c) Decomposição de Stress:

A Decomposição de Stress é usada para identificar as fontes de erro que contribuem para o Stress total do modelo. Normalmente, isso é feito considerando-se diferentes componentes de erro:

$$Stress^2 = Stress_{model}^2 + Stress_{bias}^2 + Stress_{instability}^2 + Stress_{random}^2$$

onde:

- $Stress_{model}$ é o Stress devido ao modelo de MDS escolhido;
- $Stress_{bias}$ é o Stress devido ao viés no conjunto de dados;
- $Stress_{instability}$ é o Stress devido à instabilidade nos resultados devido à dimensionalidade do espaço;
- $Stress_{random}$ é o Stress devido a variações aleatórias.

Essas fórmulas são essenciais para entender e interpretar os resultados da análise de MDS, ajudando a quantificar o quanto o modelo se desvia das distâncias reais e onde estão as principais fontes de erro ou inadequação do modelo.

Testes de permutação para MDS

A qualidade do ajuste no escalonamento multidimensional requer um conjunto abrangente de ferramentas de diagnóstico em vez de confiar somente em regras práticas de stress. Neste contexto, será aplicado o teste de permutação com vista a avaliar a bondade do ajuste do MDS.

O teste de permutação em MDS é uma técnica utilizada para avaliar a significância das configurações de MDS. Ele envolve a permutação aleatória nas matrizes de dissimilaridade e a comparação da configuração resultante com a configuração observada para determinar se a estrutura observada poderia ter ocorrido por acaso.

Os testes de permutação são métodos não paramétricos usados para avaliar a significância estatística, especialmente quando a distribuição dos dados não é conhecida. Eles envolvem a reamostragem dos dados muitas vezes para construir uma distribuição nula e comparar a estatística de teste observada com essa distribuição. As hipótese a testar no teste depermutação para MDS são:

Hipótese Nula (H_0): A configuração MDS observada poderia ser obtida por acaso a partir de uma matriz de dissimilaridade aleatória.

Hipótese Alternativa (H_1): A configuração MDS observada é significativamente diferente do que seria esperado por acaso, indicando uma estrutura real nos dados.

A nossa estatística de teste será o stress e com os valores obtidos em cada permutação formaremos a distribuiçaão nula (valores de stress assumindo hipótese nula como verdadeira).

Se o p-valor for menor do que o nível de significância de 0,05, rejeite a hipótese nula, caso contrário, não se rejeite a hipótese nula.

Cluster de Variáveis

No Escalonamento Multidimensional (MDS), um *cluster de variáveis* refere-se a um grupo de variáveis que são similares ou próximas umas das outras na configuração MDS.

Um cluster é um grupo de variáveis que são mais semelhantes entre si do que com outras variáveis. No contexto do MDS, variáveis que estão próximas no espaço MDS formam um cluster, indicando que elas têm padrões de variação semelhantes.

Para identificar os clusters, podemos recorrer a várias técnicas de agrupamento, como cluster Hierarquico. O clustering hierárquico é uma técnica de análise de agrupamento que constrói uma hierarquia de clusters. É amplamente utilizado para agrupar objetos ou variáveis com base em suas similaridades ou distâncias. Existem dois tipos principais de clustering hierárquico: aglomerativo (bottom-up) e divisivo (top-down). Aqui, vamos focar no método aglomerativo, que é o mais comum.

Clustering Hierárquico Aglomerativo

O clustering Hierárquico Aglomerativo observa os seguintes passos:

a) **Inicialização:**

Cada objeto começa em seu próprio cluster.

b) **Calcular Distâncias:**

Calcule a matriz de distâncias (ou similaridades) entre todos os pares de objetos usando uma métrica apropriada (e.g., distância euclidiana, distância de Gower).

c) **Combinar Clusters:**

Este processo consiste em encontrar dois clusters mais próximos e combinar-os em um único cluster. Existem vários métodos para determinar a proximidade entre clusters, como:

- **Single Linkage**: A menor distância entre um par de pontos em dois clusters.
- **Complete Linkag** : A maior distância entre um par de pontos em dois clusters.
- **Average Linkage**: A distância média entre todos os pares de pontos em dois clusters.
- **Ward's Method:** Minimize a soma dos quadrados das distâncias dentro dos clusters.

d) **Atualização das Distâncias:**

- Recalcula-se a matriz de distâncias considerando o novo cluster.

e) **Repetir:**

- Repita os passos c) e d) até que todos os objetos estejam combinados em um único cluster.

Dendrograma:

- O resultado é representado como um dendrograma, uma árvore hierárquica que mostra a ordem e as distâncias das combinações dos clusters.

Para comparar a qualidade de agrupamentos vamos recorrer ao estatístico *Índice do Rand Ajustado.* O Índice de Rand é definido como a soma das concordâncias entre dois agrupamentos, dividido pelo número total de pares de elementos, isto é,

$$Rand\ Index = \frac{a+b}{a+b+c+d} \quad (3.48)$$

onde:

- *a* - é o número de pares de elementos que estão no mesmo cluster nas duas partições.
- b é o número de pares de elementos que estão em clusters diferentes nas duas partições.
- C é o número de pares de elementos que estão no mesmo cluster em uma partição, mas em clusters diferentes na outra partição.
- d é o número de pares de elementos que estão em clusters diferentes em uma partição, mas no mesmo cluster na outra partição.

E o Índice de Rand Ajustado (ARI) é uma versão modificada que corrige a expectativa de agrupamentos aleatórios, que é dado por:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \quad (3.49)$$

Onde:

- RI é o Índice de Rand original.
- E[RI] é a expectativa do Índice de Rand para agrupamentos aleatórios.
- max(RI) é o valor máximo possível do Índice de Rand.

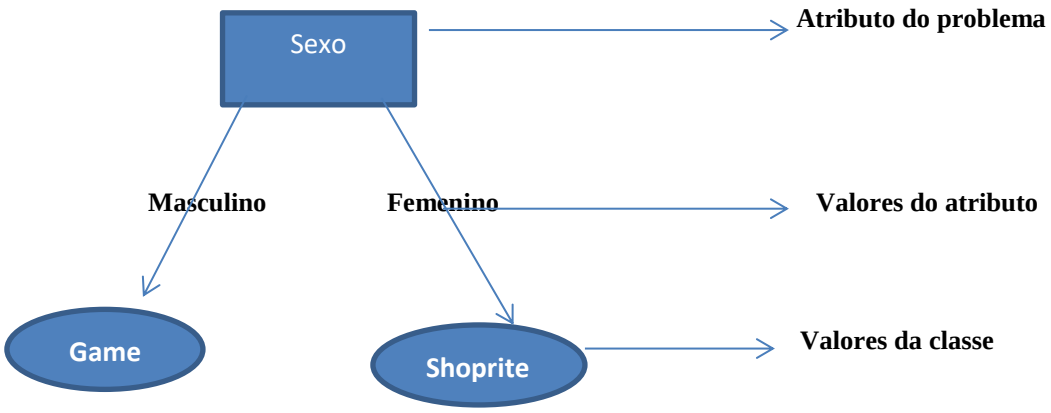
Sendo assim, vamos utilizar o estudo de estabilidade dos clusters, o que é feito obtendo várias amostras de bootstrap do conjunto de dados, aplicando o processo de cluster hierárquico e calculando o índice Rand ajustado para cada um dos agrupamentos obtidos e depois representar os resultados em diagramas de caixas para efeito de análise de estabilidade.

3.11 Árvores de decisão

Uma *árvore de decisão* (*decision tree*) é uma estrutura em forma de árvore na qual cada *nó interno* corresponde a um teste de um atributo, cada *ramo* representa um resultado do teste e os *nós folhas*

representam classes ou distribuições de classes. O nó mais elevado da árvore é conhecido como nó *raiz*, e cada caminho da raiz até um nó folha corresponde a uma *regra de classificação* (Castro & Ferrrari) .

Ilustração 2: Estrutura de árvore de classificação



Fonte: Autor

Uma vez construída a árvore, ela pode ser usada para classificar um objeto de classe desconhecida. Para isso, basta testar os valores dos atributos na árvore e percorrê-la até se atingir um nó folha, que corresponde à classe predita para aquele objeto.

Construção de árvores de decisão

A tarefa de *indução de árvores de decisão* corresponde ao processo de construção da árvore de forma que ela possa ser usada para determinar a classe de um novo objeto a partir dos valores de seus atributos. Os nós em uma árvore de decisão correspondem ao teste de determinado atributo. Em geral, o teste compara o valor do atributo a uma constante, embora algumas árvores comparem dois atributos entre si ou usem alguma função de um ou mais atributos. Os nós folhas fornecem uma classificação, um conjunto de classificações ou uma distribuição de probabilidade sobre todas as possíveis classificações, que é aplicada a todos os objetos que atingem a folha. Para classificar um novo objeto, basta apresentá-lo à raiz e caminhar na árvore até chegar a um nó folha, que dirá a classe à qual esse objeto pertence.

Se o atributo testado em um nó for nominal, então a quantidade de ramos costuma ser igual ao número de possíveis valores do atributo. Nesse caso, como há um ramo para cada valor possível, o mesmo atributo não será testado novamente na árvore.

A indução de uma árvore de decisão pode ser expressa recursivamente:

- Selecione um atributo, coloque-o na raiz da árvore e faça uma ramificação para cada valor possível, o que divide a base de dados em subconjuntos (um para cada valor do atributo);
- Repita o processo recursivamente para cada ramo, usando somente aqueles objetos que alcançam o ramo;
- Se todos os objetos em um nó possuem a mesma classificação, pare de desenvolver essa parte da árvore.

Ainda é preciso identificar qual atributo deve ser escolhido para divisão, identifica-se a quantidade de valores de cada nó, sendo que os nós com apenas uma classe, chamados *nós puros*, tornam-se folhas. Como o objetivo é encontrar árvores "económicas", quanto mais isso ocorrer, melhor. A pergunta que ainda precisa ser respondida é: Qual nó escolher para expansão? Para responder a essa pergunta, é preciso definir uma medida de *pureza* de cada nó e expandir aquele nó com filhos mais puros. A medida de pureza a ser usada é denominada *informação* e sua unidade é o *bit*. Associada a um nó da rede, a informação representa a quantidade esperada de informação que será necessária para especificar se um novo objecto deverá ser classificado em determinada classe, dado que o objecto atingiu aquele nó folha.

Cálculo da informação

Seja **X** uma base de dados com *n* objetos. Suponha que o rótulo do atributo de classe possa assumir *m* valores distintos que definem *m* classes distintas, C_i, $i = 1, ..., m$. Seja n_i a quantidade de objetos de **X** na classe C_i. A informação esperada necessária para classificar um dado objeto é:

$$I(X) = I(C_1, C_2, C_3, .., C_m) = -\sum_{i=1}^{m} p_i \log_2 p_i \qquad (3.50)$$

onde p_i é a probabilidade de que um objecto qualquer pertencer à classe C_i, estimada como n_i/n. Como normalmente os logaritmos são expressos na base 2, a unidade da informação é denominada *bits*.

Assuma que o atributo *A* possa assumir v valores distintos, $\{a_1, a_2, \ldots, a_v\}$. Ele pode ser usado para particionar **X** em *v* subconjuntos $\{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_v}\}$, onde $\boldsymbol{X_j}$ contém aqueles objectos em **X** que assumem valor a_j de *A*. Se *A* fosse seleccionado como o atributo teste, ou seja, o melhor atributo (nó) a ser expandido, então esses subconjuntos corresponderiam aos ramos que partem do nó que contém o conjunto $\boldsymbol{X_j}$.

Seja $\boldsymbol{n_{ij}}$ a quantidade de objectos da classe C_i em um subconjunto $\boldsymbol{X_j}$. A *entropia* ou *informação esperada* é dada por:

$$E(A) = \sum_{j=1}^{v} \frac{\boldsymbol{n_{1j}+\ldots+n_{mj}}}{n} . I\left(\boldsymbol{n_{1j}}+\ldots+\boldsymbol{n_{mj}}\right) \quad (3.51)$$

Onde o termo que multiplica a informação atua como um peso para o *j*-ésimo subconjunto e é o número de objectos no subconjunto dividido pelo número total de objectos em **X**.

Quanto menor o valor da entropia, maior a pureza da partição. Note que para um dado subconjunto $\boldsymbol{X_j}$ a probabilidade de um objecto em $\boldsymbol{X_j}$ pertencer à classe $\boldsymbol{C_i}$ é:

$$I\left(\boldsymbol{n_{1j}}+\ldots+\boldsymbol{n_{mj}}\right) = -\sum_{i=1}^{m} p_{ij} \log_2 p_{ij} \quad (3.52)$$

Onde $p_{ij} = n_{ij}/n_j$.

O ganho de informação a ser obtido expandindo-se *A* é:

$$ganho(A) = I(A) - E(A) \quad (3.53)$$

Em outras palavras, o ganho $ganho(A)$ é a redução esperada na entropia quando se conhece o valor do atributo *A*.

O algoritmo calcula o ganho de informação para cada atributo, e aquele com maior ganho é escolhido como o atributo teste para o conjunto **X**. Um nó é criado e rotulado com esse atributo, ramos são criados para cada valor do atributo e os objectos são particionados.

A informação é usada como base para se avaliar a expansão de um nó e deve ter as seguintes propriedades:

a) Quando o número de valores do atributo for zero, a informação é zero;
b) Quando o número de valores do atributo for igual ao número de classes, a informação possui valor máximo;
c) A informação deve obedecer a uma propriedade de múltiplos estágios.

A medida de informação está relacionada à quantidade de informação obtida ao se tomar uma decisão, sendo que uma propriedade mais útil da informação pode ser derivada considerando a natureza das decisões. As quantidades de cada um dos possíveis valores em um nó folha da árvore serão representadas na forma de uma lista de valores: $\{n_1, n_2, ..., n_m\}$, sendo que *m* é a quantidade de possíveis valores de determinado atributo.

A única função para o cálculo da informação que satisfaz as três condições listadas antes é chamada de *valor da informação* ou simplesmente *informação*

$$I(\boldsymbol{X}) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m \quad (3.54)$$

onde os termos $p_i, i, = \ 1, \ldots, n$ correspondem à probabilidade de ocorrência de cada um dos eventos *i*, $p_i = n_i/n$.

O cálculo do ganho de informação privilegia atributos com um grande número de possíveis valores. Uma forma de minimizar esse problema é empregando-se uma extensão do ganho de informação, chamada de *razão de ganho*, que aplica uma espécie de normalização ao ganho:

$$SE(A) = \sum_{j=1}^{v} \frac{n_j}{n} . \log_2 \left(\frac{n_j}{n}\right) \quad (3.55)$$

$SE(A)$ representa a informação potencial gerada dividindo-se a base de dados **X** em *v* subconjuntos correspondentes aos *v* valores possíveis do atributo *A*.

A razão de ganho é dada por:

$$Rganho(A) = \frac{ganho(A)}{SAE(A)} \quad (3.56)$$

O atributo com o maior valor da razão de ganho é seleccionado para fazer parte da árvore de decisão.

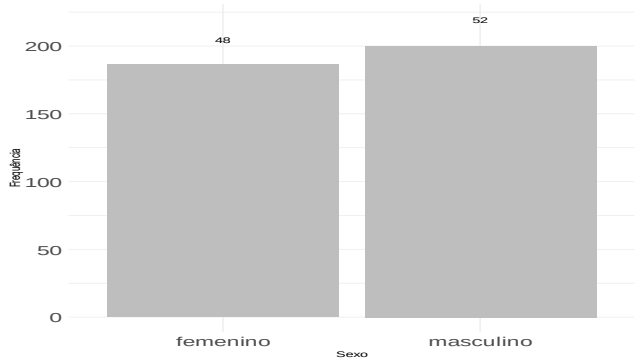
4 Capítulo IV – Apresentação e Análise de dados

4.1 Análise descritiva dos dados

Variáveis ou características do indivíduo

Foram considerados como dados ou características do individuo aquelas que recolhem os elementos que caracterizam o individuo nas suas diferentes dimensões, a saber: sexo faixa etária, estado civil e nível de escolaridade.

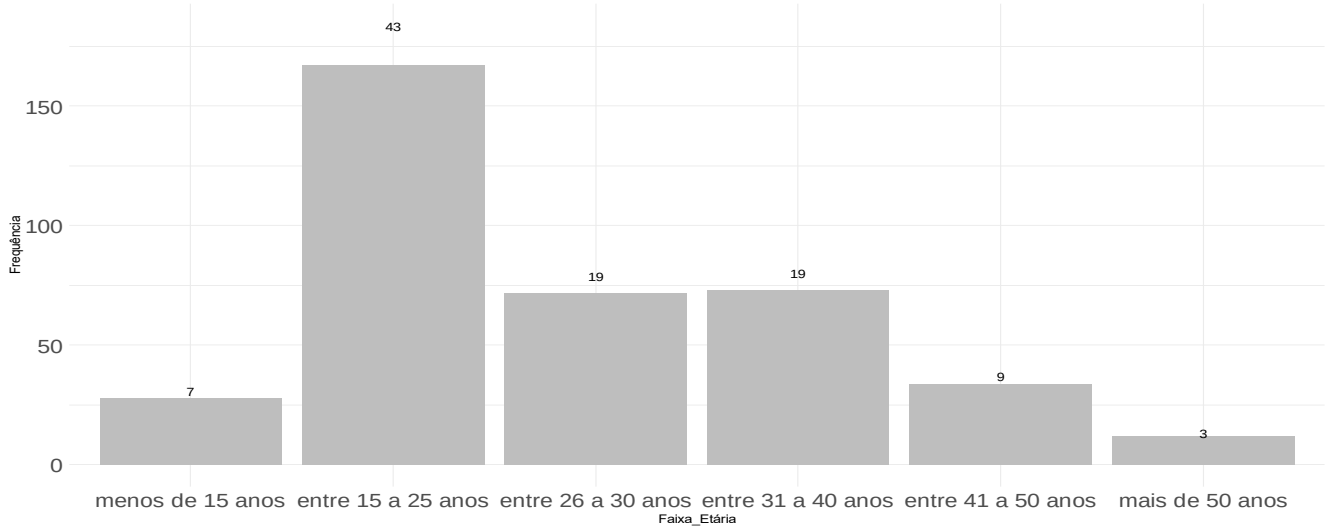
Ilustração 3: Gráfico de Distribuição dos individuos inqueridos por sexo



A nossa base de dados é composta por 386 indivíduos ou observaçõe dos quais 186 são do sexo feminino e 200 do sexo masculino, que correspondem a 48% e 52% das observações, respectivamente.

Fonte: Autor

Ilustração 4: Gráfico de Distribuição dos individuos inqueridos por faixa etária

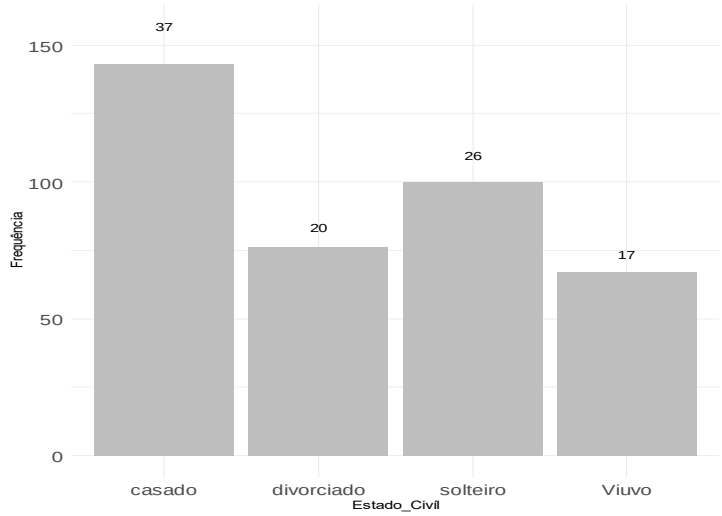


Fonte: Autor

Por faixa etária, os indivídos da nossa bese se encontram divididos em seis grupos etários, a saber:

Indivíduos cujas idades são inferiores a 15 anos em número de 28 que correspondem a 7% das observações, os com idades que variam de 15 a 25 anos, que totalizam 167 correspondente a 43%, é a faixa etária com mais observações, a seguir temos as faixas de 26 a 30 anos e 31 a 40 anos, com 72 e 73 observações respectivamente e que correspondem a 19% (as cada uma), temos a quinta faixa que varia de 41 a 50 anos com 34 observações que correspondem 9% e temos por fim, a faixa dos que têm mais de 50 anos com 12 observações que corresponde a 3%, é a faixa com menor número de observações.

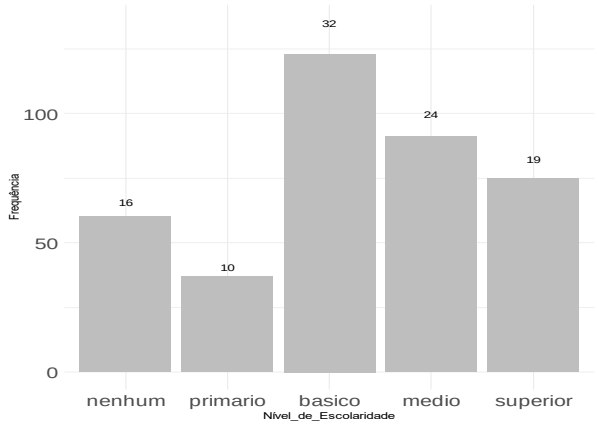
Ilustração 5: Gráfico de distribuição dos indivíduos inqueridos por estado civíl



Em termos de distribuição por estado civíl contanos com quatro grupos sendo casados em número de 143 que correspondem a 37%, é o grupo mais numeroso, divorciados em número de 73 que corresponde a 20%, temos aida o grupo de solteiros em número de 100 que corresponde a 26% e por fim, temos os viúvos num total de 67 que corresponde a 17%, este ultimo é o grupo menos numeroso.

Fonte: Autor

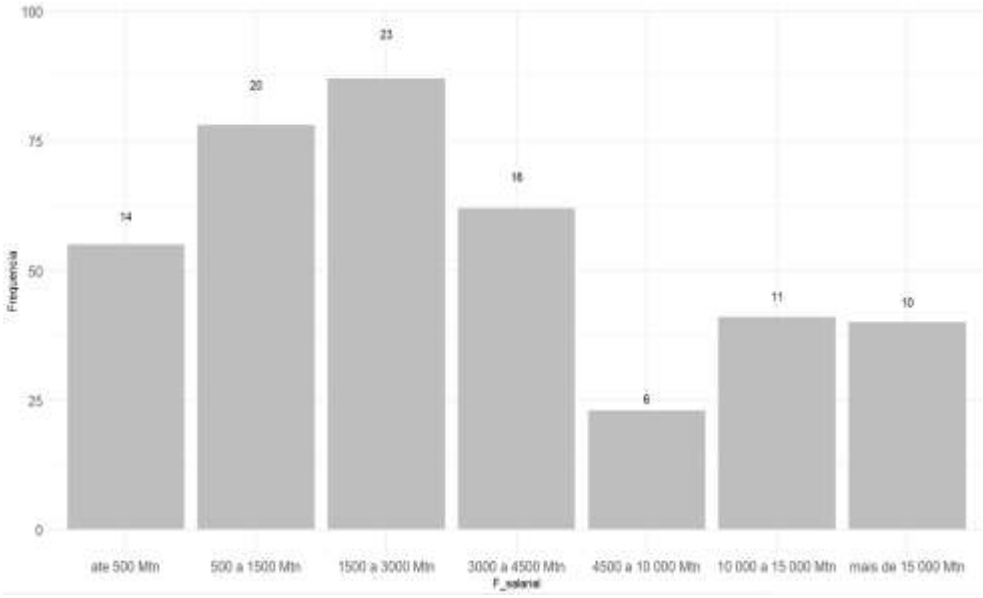
Ilustração 6: Gráfico de distribuição dos individuos inqueridos por nivel de escolaridade



Fonte: Autor

Os dados em analise apresentam-nos aida cinco grupos divididos por critério de nível de escolaridade que conta com os sem nunhum grau em número de 60 que corresponde a 16%, primário com 37 elementos correspondentes a 10%, basico com 123 elementos correspondentes a 32 elementos (é o grupo mais numeroso), nível médio com 91 elementos que correspondem a 24% e por fim, nível superior com 75 elementos correspondentes a 19%.

Ilustração 7: Gráfico de distribuição dos indivíduos inqueridos por faixa salarial



Fonte: Autor

Por fim, temos a distribuição por faixa salarial, com uma estrutura de distribuição tendente à *normal*, temos os que ganham menos de 500 Mt em numero de 55 correspondentes a 14%, de 500 a 1500 em número de 78 correspondente a 20%, de 1500 a 3000 Mt com 87 elementos correspondente a 23%, de 3000 a 4500 Mt com 62 elementos correspondente a 16%, de 4500 a 10000 Mt com 23 elementos correspondente a 6%, de 10000 a 15000 Mt com 41 elementos correspondente a 11%, e por fim, os que auferem mais de 15000 Mt com 40 elementos que corresponde a 10%.

Ilustração 8: Gráfico de distribuição dos indivíduos inqueridos por preferência de supermercado

Fonte: Autor

Do conjunto de indivíduos que responderam ao inquérito, 50% delas manifestaram-se favoráveis ao supermercado Game os outros 50% manifestaram-se favoráveis ao Shoprite. É uma boa divisão nos dados que vai favorecer na capacidade preditiva do modelo de regressão logística binária.

4.2 Aplicação da Regressão Logística Binária

Para a criação e validação dos modelos, repartiu-se a base de dados em duas partes, sendo uma de 80% correspondente a 310 observações e o resto, 20% (76 observações) para a testagem dos modelos.

Usou-se a regressão logística binária para estimar a probabilidade de um cliente ou potencial cliente dos supermercados da cidade e província de Maputo, escolher o supermercado Game ou Shoprite. Para a construção do modelo foram verificados os pressupostos da regressão logística binária com o recurso ao software (linguagem) R®. Nesta linguagem foram usados os pacotes: MASS, stargazer, car, vcd, readxl, caret, dplyr, psych e pacman. Dentro destes pacotes foram usadas as funções: {glm} usada para ajustar o modelo de regressão logística binária, {Anova} para identificar as variáveis independentes com poder explicativo para o modelo, {vif} para ver a existência de multicolinearidade, {Summary} para os coeficientes do modelo, {predic} para as previsões com dados do teste e {confusionMatrix} para achar a matriz de confusão e estimação da bondade do modelo

Pressupostos para criação do modelo de regressão logística binária

- A variável dependente Sup_preferido (supermercado de preferência) é dicotómica, isto é, tem duas categorias mutuamente exclusivas (Game ou Shoprite), o que é justificado pelo facto do questionário que deu origem à base de dados não ter dado alternativa para se optar pelas duas alternativas.
- Independência das observações, na base de dados que foi usada para este estudo não há nenhuma repetição de observações. Por isso observou-se a independência das observações.

De seguida, antes de terminar os dois ultimos pressupostos, vamos ajustar o modelo inicial de *Regressão Logistiti Binária*, de forma que, a partir deste,possaos verificar a existência de *Multicolinearidade* entre as variáveis e a existência de *pontos influientes ou outliers*, sendo o código do modelo:

ModeloInicial<-glm(Sup_preferido~., data=data_train, family = binomial (link = "logit"))

.

- Ausência de Multicolinearidade que é verificado através da *Variance inflation factor* (VIF) , se este valor for menor que 10 não há multicolinearidade (análise feita no modelo inicial), vêde a *tabela 4*, o teste revela a presença de multicolinearidade nas variáveis E_civ e N_esc (estado civíl e nível de escolaridade), testes e medidas correctivas serão realizadas de modo eliminar estas muticolinearidades.
- Não existência de pontos influentes (outliers), este pressuposto é verificado pelos resíduos padronizados onde estes devem estar entre -3 e 3. O teste com o recurso ao softwear R® sobre o modelo de *regressão logística binária (ModeloInicia)*, indica a presença de alguns outliers, mas os valores extremos são muito afastados do primeiro e terceiro quartil dos resíduos padronizados, o que pode significar que os resíduos inferiores e acima de -3 e 3, respectivamente, são insiguinificantes para comprometer o modelo, como atesta a tabela abaixo (*tabela 3)*, correspondente aos resíduos padronizados do *ModeloInicial.*

Tabela 2: Resíduos padronizados do modelo inicial

```
> summary(stdres(ModeloInicial))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-6.474552 -0.488110 -0.000664       Inf  0.547926       Inf
```

Fonte: Autor

Verificados os pressupostos preliminares (análise exploratória) sobre os dados, foi construído o modelo geral com todas variáveis independentes relacionadas com a escolha do supermercado de preferência, sendo o código R:

ModeloInicial<-glm(Sup_preferido~., data=data_train, family = binomial (link = "logit"))

que deu origem à seguinte tabela de resultados, tomando apenas tabela de saídas do :

Tabela 3: Estimação do modelo com todas variaveis

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -7.39565    1.77199  -4.174 3.00e-05 ***
SexoMasculino                    -2.44372    1.17186  -2.085 0.037038 *
F_etentre 15 a 25 anos           -0.89186    0.69648  -1.281 0.200361
F_etentre 26 a 30 anos           -0.11618    0.74964  -0.155 0.876839
F_etentre 31 a 40 anos           -0.49114    0.76261  -0.644 0.519564
F_etentre 41 a 50 anos           -0.49507    0.81527  -0.607 0.543685
F_etmais de 50 anos               0.04232    1.12362   0.038 0.969958
E_civcasado                       2.68843    0.60742   4.426 9.60e-06 ***
E_civdivorciado                   4.93337    0.89889   5.488 4.06e-08 ***
E_civviuvo                        6.19379    1.14374   5.415 6.12e-08 ***
N_escbasico                       5.35472    1.44537   3.705 0.000212 ***
N_escmedio                        0.24507    0.91770   0.267 0.789436
N_escprimario                     8.40823    1.64944   5.098 3.44e-07 ***
N_escsuperior                     3.44728    1.38304   2.493 0.012683 *
F_sal10 000 a 15 000 Mtn          3.16464    0.87521   3.616 0.000299 ***
F_sal1500 a 3000 Mtn              1.70394    0.60343   2.824 0.004747 **
F_sal3000 a 4500 Mtn              2.49519    0.64555   3.865 0.000111 ***
F_sal4500 a 10 000 Mtn            1.76748    0.90518   1.953 0.050864 .
F_sal500 a 1500 Mtn               2.34254    0.58712   3.990 6.61e-05 ***
F_salmais de 15 000 Mtn           2.25309    0.77404   2.911 0.003605 **
Tangiveismenos importante         0.23618    0.54072   0.437 0.662263
Tangiveisnao importante           0.42098    0.58977   0.714 0.475348
Prestezamais importante           0.39954    0.57469   0.695 0.486913
Prestezamais ou menos             0.67027    0.48760   1.375 0.169248
Prestezamenos importante          0.08255    0.79340   0.104 0.917136
Prestezanao importante            0.34322    0.46988   0.730 0.465126
Confiabilidademais importante     0.73980    0.66164   1.118 0.263515
Confiabilidademais ou menos      14.08966  882.74416   0.016 0.987265
Confiabilidadenao importante      0.61737    1.39133   0.444 0.657240
Segurancamais importante         -0.09789    0.69418  -0.141 0.887857
Segurancamais ou menos           -0.11483    0.49318  -0.233 0.815895
Segurancamenos importante         1.52901    1.26728   1.207 0.227613
Segurancanao e importante        -0.44219    0.52921  -0.836 0.403397
Empatiamais importante           -0.46540    0.47917  -0.971 0.331417
Empatiamais ou menos              0.10002    0.52404   0.191 0.848634
Empatiamenos importante          -0.79837    0.92516  -0.863 0.388162
Empatianao e importante           0.25126    0.43932   0.572 0.567364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 429.75  on 309  degrees of freedom
Residual deviance: 279.12  on 273  degrees of freedom
AIC: 353.12
```

Pela Tabela 2, que nos dá o efeito específico das variáveis, nota-se pelo teste de **Wald** que as variáveis *Sexo*, *E_civ*, *N_esc*, e *F_sal*, que representam o *Sexo, estado civíl, nível escolar e faixa salarial*, Respectivamente, são as que demostram alguma significância a um nível de significância de 5%. De seguida, executemos o teste VIF, para verificar a multicolinearidade das variáveis independentes.

Tabela 4: Teste de Multicolinearidade

```
> car::vif(ModeloInicial)
                    GVIF Df GVIF^(1/(2*Df))
Sexo           15.286315  1        3.909772
F_et            2.366848  5        1.089976
E_civ          29.734804  3        1.760128
N_esc         331.700464  4        2.065824
F_sal           8.139775  6        1.190925
Tangiveis       1.548656  2        1.115549
Presteza        2.522043  4        1.122585
Confiabilidade  1.865769  3        1.109540
Seguranca       2.861197  4        1.140430
Empatia         2.873340  4        1.141033
```

Fonte: Autor

Os valores de VIF das variáveis Sexo, *E_civ* e *N_esc*, apresentam um VIF elevado ($VIF > 10$), o que revela a existência de multicolinearidade entre estas variáveis e outras variáveis do modelo. Precisamos executar o teste ANOVA (do tipo II) ao modeloInicial para podermos ter os efeitos gerais do modelo e fazer a leitura das variáveis estatisticamente significativas para o modelo.

O teste do tipo II, que nos dá o efeito

Tabela 5: Teste ANOVA sobre o efeito global para do ModeloInicial

```
> Anova(ModeloInicial, type = "II")
Analysis of Deviance Table (Type II tests)

Response: Sup_preferido
               LR Chisq Df Pr(>Chisq)
Sexo              6.437  1  0.0111790 *
F_et              4.520  5  0.4771876
E_civ            47.803  3  2.345e-10 ***
N_esc            63.411  4  5.560e-13 ***
F_sal            26.432  6  0.0001849 ***
Tangiveis         0.537  2  0.7643941
Presteza          2.153  4  0.7077043
Confiabilidade    1.720  3  0.6324940
Seguranca         2.601  4  0.6266152
Empatia           3.106  4  0.5402105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

Fonte: Autor

global das variáveis, reforça o que assumimos anteriormente na tabela 5, confirmando a significância para as variaveis *Sexo, E_civ, N_esc e F_sal*. Construamos agora o novo modelo com as variáveis tidas como significativas (*modelo2).*

Adicinalmente, como forma de reforçar a credibilidade dos resultados do processo de seleção de variáveis independentes para criação do modelo, aplicamos o processo de seleção de variáveis conhecido como *stepwise* cujo critério de escolha é o menor AIC (Akaic Information Critério), tendo em conta que, Pelo teste de Critério de informação de acaique (AIC) , quanto menor for o valor melhor é o modelo e se a diferença dos valores entre os modelos não for significativa (menor ou igual a 10) escolhe-se o modelo mais simples. E o código R® é:

```
stepwise_model <- stepAIC(ModeloInicial, direction = "both")
summary(ModeloInicial)
```

e o melhor modelo é:

Tabela 6: Escolha de variáveis para o modelo óptimo pelo critério Stepwise

```
Step:  AIC=323.85
Sup_preferido ~ Sexo + E_civ + N_esc +
F_sal

                Df Deviance    AIC
<none>               293.85 323.85
- Sexo            1   298.70 326.70
+ Tangiveis       2   293.32 327.32
+ Confiabilidade  3   291.77 327.77
+ Empatia         4   290.55 328.55
+ F_et            5   288.98 328.98
+ Seguranca       4   291.48 329.48
+ Presteza        4   292.21 330.21
- F_sal           6   323.41 341.41
- E_civ           3   350.01 374.01
- N_esc           4   362.94 384.94
```

Fonte:Autor

E estes resultados reforçam a escolha de variável feita apartir doteste ANOVA pelo critério de *p-valor*

Tabela 7: Estimação do modelo 2 com apenas variaveis significativas & resíduos pagronizados

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -6.8181     1.1655  -5.850 4.91e-09 ***
SexoMasculino          -2.0172     1.1170  -1.806 0.070944 .
E_civcasado             2.6873     0.5432   4.948 7.52e-07 ***
E_civdivorciado         4.7718     0.8200   5.820 5.90e-09 ***
E_civviuvo              6.1508     1.0305   5.969 2.39e-09 ***
N_escbasico             5.1435     1.3536   3.800 0.000145 ***
N_escmedio              0.3639     0.8352   0.436 0.663042
N_escprimario           7.9830     1.5033   5.310 1.10e-07 ***
N_escsuperior           3.2830     1.3232   2.481 0.013097 *
F_sal10 000 a 15 000 Mtn   3.0659     0.7738   3.962 7.42e-05 ***
F_sal1500 a 3000 Mtn       1.5821     0.5331   2.968 0.003001 **
F_sal3000 a 4500 Mtn       2.4396     0.5932   4.113 3.91e-05 ***
F_sal4500 a 10 000 Mtn     1.7134     0.8671   1.976 0.048137 *
F_sal500 a 1500 Mtn        2.0985     0.5231   4.012 6.03e-05 ***
F_salmais de 15 000 Mtn    2.2369     0.6938   3.224 0.001264 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 429.75  on 309  degrees of freedom
Residual deviance: 293.85  on 295  degrees of freedom
AIC: 323.85

Number of Fisher Scoring iterations: 6

> summary(stdres(Modelo2))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-5.631136 -0.512911 -0.005376 -0.016928  0.572569  2.943128
```

Fonte: Autor

Agora temos todas variáveis significativas embora, algumas das suas categorias se manifestem irrelevantes, que é o caso da categoria *médio* da variável nível de escolaridade (*N_esc*).

De realçar que, ao descar as variáveis não significativas, houve melhoria na satisfação do pressuposto de multicolinearidade, vêde a tabela abaixo.

Teste de razão de verosimilhança

Com o teste de razão de verossimilhança (likelihood-ratio test), queremos verificar a adequação do ajuste do modelo completo em comparação com o modelo nulo, para o efeito usamos o seguinte código em R® :

1-pchisq(modelo2$null.deviance-modelo2$deviance, modelo2$df.null-modelo2$df.residual)

E temos como resultado zero (0), e sendo o resultado zero (menor que p-valor), rejeitamos a hipótese nula, que assume que os modelos têm a mesma capacidade preditiva

Cálculo de Odds-ratio e interpretação

Continuando, obtemos os ODDS-ratio dos efeitos do modelo, para efeito de interpretação dos parâmetros do modelo.

Tabela 8: Cálculo de chances proporcionais com intervalos de confiança

```
> exp(cbind(OR=coef(Modelo2),IC=confint.default(Modelo2)))
                                    OR        2.5 %       97.5 %
(Intercept)                 1.093762e-03 1.113980e-04 1.073912e-02
SexoMasculino               1.330289e-01 1.489773e-02 1.187878e+00
E_civcasado                 1.469176e+01 5.066831e+00 4.260014e+01
E_civdivorciado             1.181327e+02 2.368188e+01 5.892836e+02
E_civviuvo                  4.690724e+02 6.223777e+01 3.535296e+03
N_escbasico                 1.713139e+02 1.206803e+01 2.431917e+03
N_escmedio                  1.438934e+00 2.799843e-01 7.395169e+00
N_escprimario               2.930730e+03 1.539314e+02 5.579874e+04
N_escsuperior               2.665431e+01 1.992889e+00 3.564936e+02
F_sal10 000 a 15 000 Mtn    2.145383e+01 4.708342e+00 9.775561e+01
F_sal1500 a 3000 Mtn        4.865084e+00 1.711187e+00 1.383194e+01
F_sal3000 a 4500 Mtn        1.146857e+01 3.585732e+00 3.668098e+01
F_sal4500 a 10 000 Mtn      5.548016e+00 1.014142e+00 3.035126e+01
F_sal500 a 1500 Mtn         8.154271e+00 2.924823e+00 2.273373e+01
F_salmais de 15 000 Mtn     9.364030e+00 2.403826e+00 3.647729e+01
```

Fonte: Autor

Com base nesta tabela de saída de chances proporcionais, com intervalos de confiança a 95% de confiança, temos que:

Sendo o evente para este modelo, a escolha do supermercado Game como preferência e, tendo "0" como categoria de referência para a variável resposta (Sup_preferido), que corresponde ao supermercadoShuprite, temos as seguintes interpretações do modelo:

Ser do sexo masculino (*SexoMasculino)*, diminue em 87% $(OR \approx 0{,}13)$ as chances de ter o supermercado game como supermercado de preferência, em comparação com os indivíduos do sexo feminino (categoria de referência)

Ser casado (*E_civilcasado*), aumenta aproximadamente 15 vezes $(OR \approx 14{,}69)$ as chances de ter Supermercado Game como sua preferência em relação ao solteiro (categoria de referência);

Ser divorciado (E_civildivorciado), aumenta 118 vezes $(OR \approx 118{,}13)$ as chances de ter supermercado Game como sua preferência em relação ao solteiro;

Ser viúvo (E_civilviuvo), aumenta em 469 vezes $(OR \approx 469{,}7)$ as chances de ter Supermercado Game como sua preferência em relação ao solteiro.

Tendo nível de escolaridade primário (*N_escolarprimario*) , as chances de ter supermercado Game como preferência aumentam 2930 veses $(OR \approx 2930{,}7)$, em relação a quem não tem nenhum nível de escolaridade;

Tendo nível de escolaridade básico (*N_escbasico*) , as chances de ter supermercado Game como preferência aumentam 171 veses ($OR \approx 171{,}31$) , em relação a quem não tem nenhum nível de escolaridade;

Tendo nível superior de escolaridade (*N_escolarsuperior*), as chances de ter supermercado Game como preferência aumenta aproximadamente 27 vezes ($OR \approx 26{,}65$), em relação a quem não tem nenhum nível;

Ganhar entre 500 e 1500 (*F_salarial500 a 1500 Mtn*) aumenta aproximadamente 8 vezes mais ($OR \approx 8{,}15$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn

Ganhar entre 1500 a 3000 (*F_salarial1500 a 3000 Mtn*) aumenta aproximadamente 4 vezes mais ($OR \approx 4{,}865$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn;

Ganhar entre 3000 a 4500 (*F_salarial3000 a 4500 Mtn*) aumenta aproximadamente 11 vezes mais ($OR \approx 11{,}47$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn;

Ganhar entre 4500 a 10000 (*F_salarial4500 a 10000 Mtn*) aumenta aproximadamente 6 vezes mais ($OR \approx 5{,}55$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn;

Ganhar entre 10000 a 15000 (F_salarial10000 a 15000 Mtn) aumenta aproximadamente 21 vezes mais ($OR \approx 21{,}45$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn;

Ganhar mais de 15000 Mtn (*F_salarialmais de 15000 Mtn*) aumenta aproximadamente 9 vezes mais ($OR \approx 9{,}36$) as chances de ter supermercado Game como sua preferência, em relação a que ganha menos que 500 Mtn.

Avalição do Modelo

Matriz confusão, Acurácia, Especificidade e Sensitividade

Segue neste momento, com recurso às ferramentas do software R® (biblioteca caret) a avaliação do modelo discutido no parágrafo anterior nos seus diferentes aspectos ou medidas a começar pela *matriz*

de confusão, que requer as previsões dadas pelo modelo sobre as observações usadas na construção do mesmo, para tal vamos trabalhar com um *cutoff* de 0,5 para a classificação das observações que são evento e as não evento , o que é um bom *cutoff* dado que, as observações na base de dados são 50% evento, e classificaremos como evento aquelas observações cuja probabilidade é igual ou superior a 0,5.

Tabela 9: Matriz confusão, Acurácia, sensitividade e especificidade do Modelo2

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 117  32
         1  38 123

               Accuracy : 0.7742
                 95% CI : (0.7235, 0.8195)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5484

 Mcnemar's Test P-Value : 0.5501

            Sensitivity : 0.7935
            Specificity : 0.7548
         Pos Pred Value : 0.7640
         Neg Pred Value : 0.7852
             Prevalence : 0.5000
         Detection Rate : 0.3968
   Detection Prevalence : 0.5194
      Balanced Accuracy : 0.7742

       'Positive' Class : 1
```

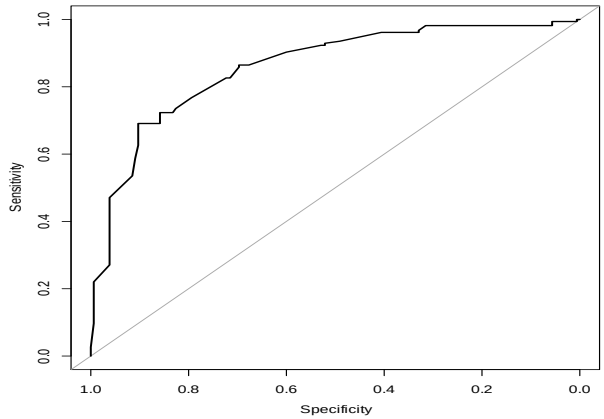
Fonte: Autor

O Modelo2 consegue classificar correctamente 77% dos dados e dá origem a um índice Kappa de 0,5484 (moderado) o que indica uma concordância em quase 60% das previsões do modelo com as observações. Além disso, podemos ver como o modelo consegue detectar “eventos” e “não eventos”, pois a sensibilidade (percentual de acertos considerando-se apenas as observações que de facto são evento) e a especificidade (percentual de acerto considerando-se apenas as observações que não são evento) que assumem valores diferentes 79,35% e 75,48% respectivamente.

4.2.1.1 Curva Roc

De modo a ter uma outra perspectiva sobre a eficiência global do modelo, construamamos a curva de Roc.

Ilustração 9: Curva Roc do Modelo2 para os dados de treinamento



Fonte: Autor

A curva de Roc apresenta uma área sobre ela (AUC) correspondente a 0,8566 (próximo de 1) o que revela boa eficiência global do modelo

De modo a avaliar a capacidade preditiva do Modelo2, usamos agora a partição de teste (data_test) para obter estimativas mais precisas.

Tabela 10: Ilustração 10: Matriz confusão, Acurácia, sensitividade e especificidade do Modelo2 (dados do teste)

```
          Reference
Prediction  0  1
         0 30  8
         1  8 30
```

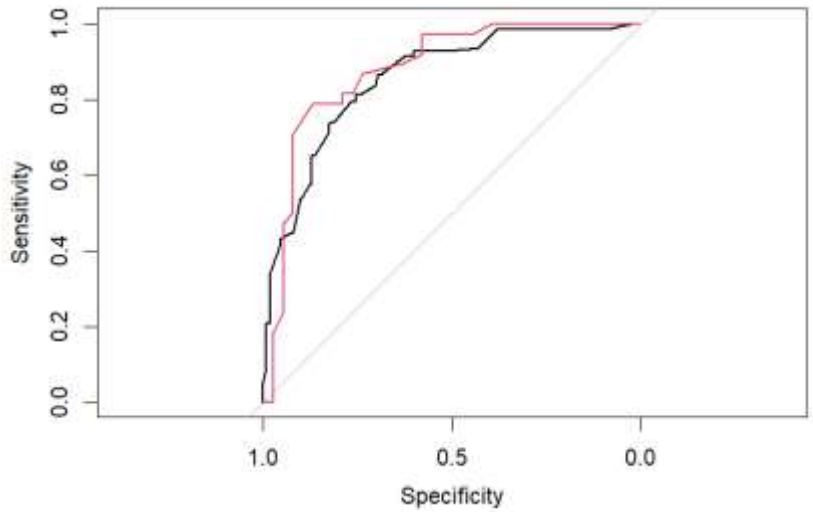
```
               Accuracy : 0.7895
                 95% CI : (0.6808, 0.8746)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.925e-07

                  Kappa : 0.5789

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7895
            Specificity : 0.7895
         Pos Pred Value : 0.7895
         Neg Pred Value : 0.7895
             Prevalence : 0.5000
         Detection Rate : 0.3947
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.7895

       'Positive' Class : 1
```

Fonte: Autor

Nota-se um ligeiro melhoramento dos parâmetros indicadores da qualidade do modelo, em quase 2%, em quase todos eles, com a excepção da especificidade que que subiu em quase 4% enquanto que a sensitividade não teve nenhuma alteração significativa, o que revela uma relativa estabilidade do modelo

Construamos agora as curva ROC e das áreas (AUC) com base nos dados de treinamento e teste, representando no mesmo sistema, sendo treinamento (preto) e teste (vermelho) para fácil comparação:

Ilustração 10: Curva Roc do Modelo2 (com dados de treinamento e teste, a preto e vermelho respetivamente)



Fonte: Autor

Nota-se um ligeir aumento do valor da área sobre a curva Roc de teste (0,8792), em comparação com o de treinamento que é de 0,8366. Este ligeiro aumento reforça a ideia de estabilidade do modelo.

Conclusão sobre o modelo de Regressão Logística Binária

Conforme o que os testes vieram a revelar, desde os testes de adequação dos dados para a criação do modelo de regressão logística binária (pressupostos), passando pelos testes de significância das variáveis independentes e adequação do modelo, podemos concluir que o modelo que melhor prevê a probabilidade de um cliente ou potencial cliente optar pelo supermercado Game ou Shoprite como supermercado da sua preferência é o Modelo2 cuja representação é:

$$P(Sup\_preferido = Game) = \frac{1}{1+e^{-Z}},$$

Onde: $Z = 0{,}109 + 0{,}133 * SexoMasculinor +$

$$+14{,}691 * E\_civCasado + 118{,}133 * E\_civDivorciado + 469{,}072 * E\_civViuvo +$$

$$+171{,}314 * N\_escBasico + 2930{,}73 * N\_escPrimario + 26{,}654 * N\_escSuperior +$$

$$+8{,}154 * F\_sal500a1500 + 4{,}865 * F\_sal1500a3000 + 11{,}469 * F\_sal3000a4500 +$$

$$+5{,}548 * F\_sal4500a10000 + 21{,}454 * F\_sal10000a15000 + 9{,}364 * F\_salmaisde15000$$

4.3 Aplicação da Regressão Logística Ordinal

No presente parágrafo iremos realizar um estudo sobre a base de dados *satisfação_com_o_supermercado_game_ou_shoprite,* este estudo consiste em encontrar os elementos que essencialmente caracterizam os indivíduos que alcançam os diferentes níveis de satisfação, encontrar um modelo que permite determinar a probabilidade de um determinado cliente (observação) alcançar certo níveil de satisfação junto aos supermercados Game e Shoprite da provincia e cidade de Maputo e analizar até que ponto o modelo ora criado pode ser eficiente. Os níveis de satisfação que são as categorias da variável resposta, ordenados de 1 a 5 são: pouco satisfeito, pouco, mais ou meno, muito e muito satisfeito. E estas categorias estão distribuidas na base conforme as proporção abaixo:

```
 pouco satisfeito     pouco         mais ou menos    muito          muito satisfeito
0.06476684           0.13730570    0.12435233       0.34974093     0.32383420
```

Verificação dos pressupostos para o ajuste do modelo

Antes de ajustar o modelo vamos verificar os pressupostos para aplicação do modelo de regressão logística ordinal para a base de dados em estudo, que são:

I. O primeiro pressuposto é o *carácter ordinal* da variável dependente. A nossa variável dependente é *nivel de satisfação* cujas categorias expressam os diferentes níveis de satisfação, o que subentende a presença de ordenação nas categorias.

II. O segundo pressuposto tem a ver com a *independência* das observacões, e para verificar este pressuposto recorremos ao teste de Durbin que é uma extensão do teste de Durbin-Watson para modelos de regressão logística ordinal, que avalia se há autocorrelação nas diferenças de respostas ordenadas entre as observações, que é realizado sobre os resíduos generalizados do modelo, que varia de zero a quatro, para *DW=2* entende-se como ausência de corelações e um desvio considervel indica a presença de outocorrelações que podem ser negativos ou positivas conforme o valor de esttístico *DW* seja maior ou menor que 2 respectivamente. O resultado do teste indica se há evidências de autocorrelação nos resíduos.

    Para o nosso teste, sendo o *p-valor* associado ao teste maior que o nível de significância de 0,05, ou indo pelo valor do estatístico *DW* que é próximo de 2 (1,9122), podemos concluir que não há

evidências estatísticas suficientes para rejeitar a hipótese nula de independência dos resíduos. E o resultado do teste é:

```
        Durbin-Watson test

data:  residuos ~ 1
DW = 1.9122, p-value = 0.1936
alternative hypothesis: true autocorrelation is greater than 0
```

III. O terceiro pressuposto tem a ver com a *multicolinearidade*, isto é, a presença de correlações altas entre as variáveis independentes. Sendo assim, recorrendo ao teste VIF (Variance Inflation Factor) sobre o modelo para verificar este pressuposto, que considera como correlação alta quando o valor VIF for superior a 10. Sendo assim, primeiramente vamos ajustar o modelo de regressão logistica ordinal usando a função *polr* do pacote MASS, cujo código R® é:

Modelo: modelo<-polr(Nivel_satisfacao ~ ., data=data_train, start=aux)
De seguida, executamos o teste VIF e o resultastado é:

Tabela 11: Saída do estatístico VIF do Modelo

```
> vif(modelo)
      Sexo       F_et      E_civ      N_esc      F_sal Sup_preferido Tempo_cliente
  1.055961   1.110233   1.146966   1.303035   2.222209     1.030111     1.067271
  Freq_sup Outro_sup Importancia_sup  Tangiveis Confiabilidade   Presteza   Seguranca
  1.152749   1.102578       2.052924   4.766348       3.457282  12.488635   7.113251
    Empatia
 12.698769
```

Fonte: Autor

Observando os resultados do teste VIF, podemos constatar que com excepção das variáveis *Presteza* e *Empatia*, todas variáveis tem um VIF muito abaixo de 10 (menor que 5), sendo que podemos avançar com o ajuste do modelo mas prestando atenção ao facto destas variáveis com VIF alto não serem fidedignas na interpretação dos seus coeficientes.

IV. O quarto pressuposto é a presença de *chances proporcionais* (Proportional Odds) cujas hipótese são:

H_0: as chances são proporcionais ($p>0,05$)

H_1: as chances não são proporcionais ($p<0,05$)

E o código R® : car::poTest(Modelo)

Tabela 12: Teste de chances proporcionais

	Pr(>Chisq)
Overall	1
SexoMasculino	1
E_civcasado	1
E_civdivorciado	1
E_civviuvo	1
N_escbasico	1
N_escmedio	1
N_escprimario	1
N_escsuperior	1
F_sal10 000 a 15 000 Mtn	1
F_sal1500 a 3000 Mtn	1
F_sal3000 a 4500 Mtn	1
F_sal4500 a 10 000 Mtn	1
F_sal500 a 1500 Mtn	1
F_salmais de 15 000 Mtn	1
Outro_supLM	1
Outro_supLuz	1
Outro_supMohmed e Companhia	1
Outro_supNosso Supermercado	1
Outro_supNovo Mundo	1
Outro_supOutro	1
Confiabilidademais importante	1
Confiabilidademais ou menos	1
Confiabilidadenao importante	1

Fonte: Autor

Como se pode observar, através dos p_valores, tanto a variável depenndente (modelo) como as variáveis regressoras, mostram a existência de chances proporcionais. O que torna numa prática correcta a criação, análise e interpretação do modelo de regressão logística ordinal.

Com base no código do modelo acima indicado, temos a seguir as saidas do modelo com os respetivos valores de pontos de corte (α_i) ou interceptos.

Tabela 13: Coeficientes de modelo de RLO

	Value	Std. Error	t value	p value
SexoMasculino	-31.6468	5.4029	-5.857400e+00	0.0000
E_civcasado	-9.2925	31.6288	-2.938000e-01	0.7689
E_civdivorciado	-27.0515	5.4380	-4.974500e+00	0.0000
E_civviuvo	-29.9012	5.3309	-5.609000e+00	0.0000
N_escbasico	3.8887	1.7481	2.224500e+00	0.0261
N_escmedio	6.3124	0.0000	1.657488e+05	0.0000
N_escprimario	5.5710	1.5254	3.652200e+00	0.0003
N_escsuperior	-0.9296	1.4983	-6.205000e-01	0.5349
F_sal10 000 a 15 000 Mtn	0.8916	1.2906	6.908000e-01	0.4897
F_sal1500 a 3000 Mtn	0.2255	0.8564	2.634000e-01	0.7923
F_sal3000 a 4500 Mtn	3.1985	1.1369	2.813300e+00	0.0049
F_sal4500 a 10 000 Mtn	3.6592	1.2384	2.954700e+00	0.0031
F_sal500 a 1500 Mtn	1.8226	0.8416	2.165700e+00	0.0303
F_salmais de 15 000 Mtn	0.2027	1.2468	1.625000e-01	0.8709
Outro_supLM	-8.4211	1.5888	-5.300100e+00	0.0000
Outro_supLuz	-17.7537	26.5122	-6.696000e-01	0.5031
Outro_supMohmed e Companhia	-2.3426	1.2659	-1.850500e+00	0.0642
Outro_supNosso Supermercado	-21.8229	26.5134	-8.231000e-01	0.4105
Outro_supNovo Mundo	14.9123	0.0005	3.005668e+04	0.0000
Outro_supOutro	-27.3948	0.0000	-8.586604e+09	0.0000
Confiabilidademais importante	-1.9620	0.9510	-2.063200e+00	0.0391
Confiabilidademais ou menos	-11.1813	40.3736	-2.769000e-01	0.7818
Confiabilidadenao importante	-3.1968	0.0000	-2.365220e+05	0.0000
pouco satisfeito|pouco	-63.2056	10.6056	-5.959600e+00	0.0000
pouco|mais ou menos	-46.4895	36.8207	-1.262600e+00	0.2067
mais ou menos|muito	-34.1255	5.4128	-6.304600e+00	0.0000
muito|muito satisfeito	-25.8135	5.3887	-4.790300e+00	0.0000

Fonte: Autor

Observando a tabela de saída do modelo, pode se perceber que temos todas variáveis do modelo significativas. É de notar tambem que o modelo não considera relevante a transição da categoria *pouco* para *mais ou menos*, itso é, o modelo não considera estas categorias como de níveis diferentes.

Análise do tipo II (teste de Wald)

De modo a obter o efeito global das variáveis, vamos executar o teste ou análise de tipo II, de modo a identificar as variáveis significativas no modelo inicial, e construair um novo modelo com as variáveis tidas como significativas.

Tabela 14: Resultados do teste do tipo II

```
typeII
              var     LRChisq Df          pVal
            Sexo  63.1311257  1 1.887379e-15
            F_et   4.1329036  5 5.304442e-01
           E_civ 308.8012464  3 0.000000e+00
           N_esc   7.8624402  4 9.674916e-02
           F_sal  20.0507177  6 2.712411e-03
   Sup_preferido   2.5008689  1 1.137835e-01
   Tempo_cliente   4.1859356  3 2.420738e-01
        Freq_sup   6.9124235  4 1.405893e-01
       Outro_sup  73.2896869  6 8.626433e-14
 Importancia_sup   3.8046703  4 4.330858e-01
       Tangiveis   0.5648484  2 7.539538e-01
        Presteza   3.9825301  4 4.083753e-01
  Confiabilidade   9.8637107  3 1.976089e-02
       Seguranca   3.8085041  4 4.325419e-01
         Empatia   1.3559433  4 8.518150e-01
```

Fonte: Autor

Observando as saidas do teste do tipo II ao lado, percebe-se que as variáveis Sexo, estado civil, faixa salarial, outro supermercado e confiabilidade, são as que demostram alguma sigmificância estatística para a predição da variável depenente. Sendo assim, vamos de seguida ajustar um novo modelo com estas variáveis.

Teste de razão de verosimilhança

Com o teste de razão de verossimilhança (likelihood-ratio test), queremos verificar a adequação do ajuste do modelo completo em comparação com o modelo nulo, para o efeito usamos o seguinte código em R®:

1-pchisq(Modelo$null.deviance-Modelo$deviance, Modelo$df.null-Modelo$df.residual)

E temos como resultado zero (0), e sendo o resultado zero (menor que p-valor), rejeitamos a hipótese nula, que assume que os modelos têm a mesma capacidade preditiva.

Cálculo de Odds-ratio

De seguida, vamos determinar os Odds-ratio dos efeitos do modelo, para efeito de interpretação dos parâmetros do modelo de regressão logística Ordinal.

Tabela 15: Chances proporcionais e intervalo de confiaça

	OR	2.5 %	97.5 %
SexoMasculino	1.802828e-14	4.539602e-19	7.159635e-10
E_civcasado	9.211099e-05	1.101188e-31	7.704799e+22
E_civdivorciado	1.785267e-12	4.196132e-17	7.595517e-08
E_civviuvo	1.032935e-13	2.995084e-18	3.562350e-09
N_escbasico	4.884599e+01	1.587986e+00	1.502488e+03
N_escmedio	5.513731e+02	5.513320e+02	5.514143e+02
N_escprimario	2.626927e+02	1.321461e+01	5.222059e+03
N_escsuperior	3.946955e-01	2.093761e-02	7.440416e+00
F_sal10 000 a 15 000 Mtn	2.438986e+00	1.943809e-01	3.060306e+01
F_sal1500 a 3000 Mtn	1.252982e+00	2.338834e-01	6.712587e+00
F_sal3000 a 4500 Mtn	2.449623e+01	2.638483e+00	2.274282e+02
F_sal4500 a 10 000 Mtn	3.882907e+01	3.427773e+00	4.398473e+02
F_sal500 a 1500 Mtn	6.188069e+00	1.189066e+00	3.220358e+01
F_salmais de 15 000 Mtn	1.224663e+00	1.063509e-01	1.410238e+01
Outro_supLM	2.201825e-04	9.780437e-06	4.956868e-03
Outro_supLuz	1.948255e-08	5.277035e-31	7.192857e+14
Outro_supMohmed e Companhia	9.607995e-02	8.037072e-03	1.148597e+00
Outro_supNosso Supermercado	3.329859e-10	8.999084e-33	1.232121e+13
Outro_supNovo Mundo	2.994487e+06	2.991577e+06	2.997401e+06
Outro_supOutro	1.266500e-12	1.266500e-12	1.266500e-12
Confiabilidademais importante	1.405719e-01	2.179826e-02	9.065150e-01
Confiabilidademais ou menos	1.393187e-05	5.997390e-40	3.236356e+29
Confiabilidadenao importante	4.089150e-02	4.089042e-02	4.089259e-02

Fonte: Autor

Com base nesta tabela, que nos apresenta os odds ratio, e os respectivos intervalos de confiança, cabe fazer as seguintes interpretações:

Sendo o sexo feminino a categoria de referencia para a variável *Sexo*, o sexo masculino com *OR* muito abaixo de zero, podemos afirmar que os indivíduos do sexo masculino têm muito menos chances (muito proximo de zero) de alcançar níveis mais altos de satisfação que as do sexo feminino, o mesmo é válido para indivíduos divorciados e viúvos em relação aos solteiros;

Já para o nível de escolaridade, sendo a categoria de referencia *nenhum* nível, os indivíduos com nível de escolaridade primário, básico e médio têm 262,7, 48,9 e 551,4 vezes mais chances de alcançar níveis de satisfação mais altos do que os que não têm nenhum nível escolar, respectivamente, ressalvar que a categoria *nível superior* não foi tida como relevante nos testes de significância.

Os indivíduos cujas faixas salariais variam de 500 a 1500 Mtn, 3000 a 4500 Mtn, 4500 a 10 000 Mtn, têm 6, 24,5 e 38,8 vezes mais chances de alcançar níveis de satisfação mais altos do que os que ganham abaixo de 500 Mtn, respectivante;

Em termos de fruequência de outros supermercados, os que frequentam os supermercados: LM, Nosso Supermercado e Outro Supermercado têm quase chance nula de alcançar altos níveis de satisfação em comparação aos que frequentam o supermercado Ganha Pouco (referência), enquanto que, os que frequentam os supermercados: Mohmed Companhia e Novo Mundo, têm quase 90% de chances a menos e quase 3000000 vezes mais de chances, respectivamente, de alcançar níveis de satisfação mais altos do que os que frequentam o supermercado Ganha Pouco.

Em termos da variável *Confiabilidade*, os que assumen este atributo como *mais importante* e *não importante*, têm 86% e 96% de chances a menos de alcançar níveis de satisfação mais altos em comparação com os que consideram este atributo importante, respectivamente.

Avalição do Modelo

Depois da construção e interpretação do modelo, segue neste momento, com recurso às ferramentas do software R® (biblioteca caret) a avaliação do modelo discutido no parágrafo anterior nos seus diferentes aspectos ou medidas a começar pela *matriz de confusão*, que requer as previsões dadas pelo modelo sobre as observações usadas na construção do mesmo.

Tabela 16: Matriz confusão, índice Kappa e acurácia (dodos de treinamento)

	pouco satisfeito	pouco	mais ou menos	muito	muito satisfeito
pouco satisfeito	18	1	0	0	0
pouco	2	40	4	0	0
mais ou menos	0	2	35	0	2
muito	0	0	0	107	5
muito satisfeito	0	0	0	1	93

Accuracy	Kappa
0.9451613	0.9254860

Fonte: Autor

Observando a matriz confusão, podemos perceber que há pouca ocorrência dos erros, isso nota-se observando os valores na diagonal principal e fora desta na matriz principal. Esta conclusão é reforçada pelos valores de Acurácia (0,9451613) que significa que o modelo classifica correctamente 94,5% das observações e índice de Kappa (0,9254860) que significa que a concordância entre as previsões do modelo e os dados observados é quase perfeita.

Tabela 17: Índice Kappa ponderado (dados de treinamento)

	value	ASE	z	Pr(>|z|)
Unweighted	0.9255	0.01745	53.04	0
Weighted	0.9536	0.01144	83.36	0

Weights:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.00	0.75	0.50	0.25	0.00
[2,]	0.75	1.00	0.75	0.50	0.25
[3,]	0.50	0.75	1.00	0.75	0.50
[4,]	0.25	0.50	0.75	1.00	0.75
[5,]	0.00	0.25	0.50	0.75	1.00

Fonte: Autor

Observando a matriz pode se concluir que:

Equivocar-se um nível corresponde a 75% de acerto, dois níveis 50% de acerto e 3 níveis 25% de acerto.

Obtem-se um índice de Kappa ponderado de 0,9535 contra um índice Kappa clássico de 0,9255, cujo contraste ou teste de hipótese mostra uma sigificância estatística diferente de zero, sustentando a hipótese de que os mesmos dão melhor classificação do que se obteria pelo acaso.

Para obtermos estatísticas mais realistas usamos agora a partição dos dados correspondente ao teste (data_teste).

Com o código:

cm<-confusionMatrix(table(predict(Modelo, newdata=data_test),data_test$Nivel_satisfacao)),

obtemos a matriz confução:

Tabela 18: Matriz confusão, índice Kappa e acurácia (dodos deteste)

	pouco satisfeito	pouco	mais ou menos	muito	muito satisfeito
pouco satisfeito	5	1	0	0	0
pouco	0	9	1	0	0
mais ou menos	0	0	8	0	0
muito	0	0	0	27	2
muito satisfeito	0	0	0	0	23

Accuracy	Kappa
0.9473684	0.9278937

Fonte: Autor

Que tambem não apresenta de forma significativa uma extrutura diferente da matriz construida com dados de treinamento.

E tambem não altera praticamente o valor da acurácia de 0,9473684 contra 0,9451613 obtido apartir dos dados e treinamento, de modo semelhante o índice de Kappa tambem não regista nenhuma varição significativa sendo de 0,9278937 contra o valor 0.9254860 obtido a partir dos dados de treinamento.

Quanto ao kappa ponderado e a sua significância, mantêm-se os valores obtidos a partir dos dados de treinamento, e continuam observando o mesmo sentido de significância estatística.

Tabela 19: Índice Kappa ponderado (dados de teste)

	value	ASE	z	Pr(>|z|)
Unweighted	0.9279	0.03491	26.58	1.101e-155
Weighted	0.9602	0.01932	49.70	0.000e+00

Weights:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.00	0.75	0.50	0.25	0.00
[2,]	0.75	1.00	0.75	0.50	0.25
[3,]	0.50	0.75	1.00	0.75	0.50
[4,]	0.25	0.50	0.75	1.00	0.75
[5,]	0.00	0.25	0.50	0.75	1.00

Fonte: Autor

Em suma, o modelo de regressão logística Ordinal mantem o seu desempenho quando é submetido aos dados de teste, sendo assim podemos afirmar que o mesmo é robusto.

4.4 Árvores de Classificação

Árvore de Classificação para Modelo de Classificação Binária

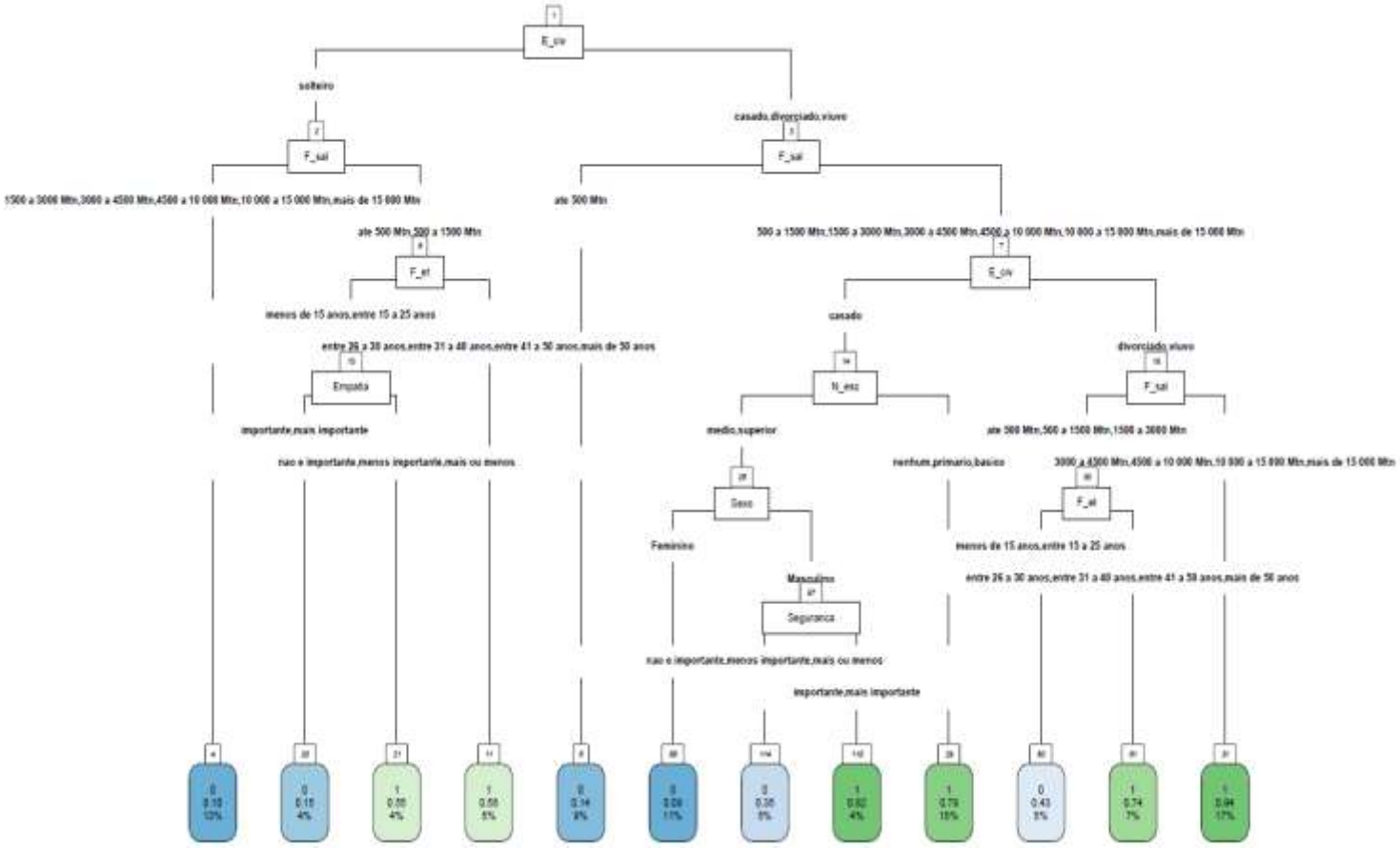
No presente parágrafo queremos criar um modelo que permite classificar os inqueridos em dois grupos, os favoráveis ao supermercao *Game* ou *Shoprite*, em alternativa ao modelo de *Regressão logística binária.* Para esse feito vamos usar a funão *rpart* do pacote R®. À semelhança do modelo de regressão logística binária, vamos trabalhar com as variáveis : carecterísticas do indivíduo (sexo, faixa etária, estado civíl, faixa salarial e estado civíl), ordem de valores (Empatia, Segurança, Presteza , Confiabilidade e Tangíveis) e a própria variável resposta que é supermercado de preferência.

Os dados são repartidos em duas partes, uma para treinamento (80%) e outra para prova ou teste (20%).

a) Criação e interpretação da árvore

Para criação da árvore de classificação, em conformidade com a literatura usada nesta pesquisa, vamos usar o critério de *Entropia.* Com o recurso à função *rpart* do pacote do mesmo nome do software R®, construimos a árvore de classificação a seguir:

Ilustração 11: Árvore de classificação binária (maximal)



Fonte: Autor

Construida a árvore de classificação pelo critério de entropia, vamos agora estudar a eficiência da mesma em termos de capacidade preditiva. Para a materialização desse facto vamos costruir a matriz de confusão ou classificação, tomando como cutoff de 50%, conforme a persentagem dos eventos nos dados (observações). Executando os comandos R® para o cálculo da matriz temos as saidas:

Tabela 20: Matriz cofusão, acurácia, índice de kappa, sensitividade e especificidade

```
> cm$table
          Reference
Prediction  0  1
         0 33  9
         1  5 29
```

Accuracy	Kappa
0.8157895	0.6315789

Sensitivity	Specificity
0.7631579	0.8684211

Fonte: Autor

Como se pode observar acima, o nosso modelo (árvore), tem aproximadamente 82% de acurácia, isto é, a taxa de classificação correcta de dados é de 82%, Possue também um indice de Kappa aproximadamen

e igual a 0,63 que entende-se como concordância substancial (entre as previsões e as observações). Te mos sensitividade aproximadamente igual a 0,76, que entende-se como capacidade do modelo detectar v erdadeiros positivos, que neste caso é boa. Temos tambem uma especificidade de aproximadamente igua l a 0,87, que é a capacidade do modelo detectar correctamente os verdadeiros negativos, que neste caso t ambem é aceitável.

b) Importância das Variáveis

De seguida, vamos analizar a importância das variáveis de modo a aferir quais são as variáveis que dão mais informações na hora de criação do modelo (árvore). Para esse feito vamos executar os comandos de R®,

modeloEntropia$variable.importance

*modeloEntropia$variable.importance/sum(modeloEntropia$variable.importance)*100*

que dão o resuldo abaixo:

Tabela 21: Importância de variáveis

F_sal	E_civ	N_esc	Sexo	Seguranca	F_et	Empatia
33.451104	29.204172	17.703705	9.731588	3.814388	3.477754	2.617288

Fonte: Autor

Portanto, a variável mais importante é a faixa salarial (*F_sal* com 33,5%, seguido de estado civíl (*E_civ)* com 29,2%, nível de escolaridade, com 17,7%, sexo com 9,7%, segurança 3,8%, faixa etária com 3,5% e empatia com 2,6% de importância.

De seguida temos a representação gráfica da importância das variáveis.

Gráfico 10: Importância de Variáveis

Ilustração 12: Grafico da Importância de variáveis

Fonte: Autor

Poda da árvore

Com o objectivo de reduzir o super ajuste do modelo (árvore), neste parágrafo vamos execur a poda da árvore construida anteriormente, de modo a obter uma árvore com menos folhas (mais parcimoniosa), se esta for a melhor alternativa.

De seguida vamos construir a *tabela de complexidade* - *cp* que nos permite visualizar as diferentes subarvores e os respectivos erros associados, para esse efeito vamos executar o comando R® abaixo.

modeloEntropia$cptable

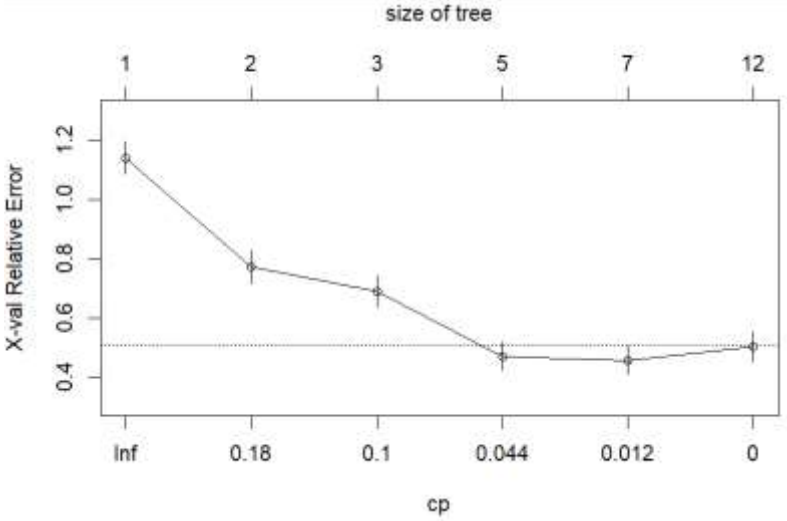
Tabela 22: Tabela de complecidades

```
> modeloEntropia$cptable
           CP nsplit rel error    xerror       xstd
1 0.251612903      0 1.0000000 1.1419355 0.05622117
2 0.129032258      1 0.7483871 0.7741935 0.05532926
3 0.083870968      2 0.6193548 0.6903226 0.05400418
4 0.022580645      4 0.4516129 0.4709677 0.04819734
5 0.006451613      6 0.4064516 0.4580645 0.04773265
6 0.000000000     11 0.3741935 0.5032258 0.04929226
```

Fonte: Autor

Com esta tabela de complexidade podemos escolhher a subárvore que minimiza o erro e overfitting, mas a melhor alternativa podemos obter a partir do gráfico de complexidade a seguir.

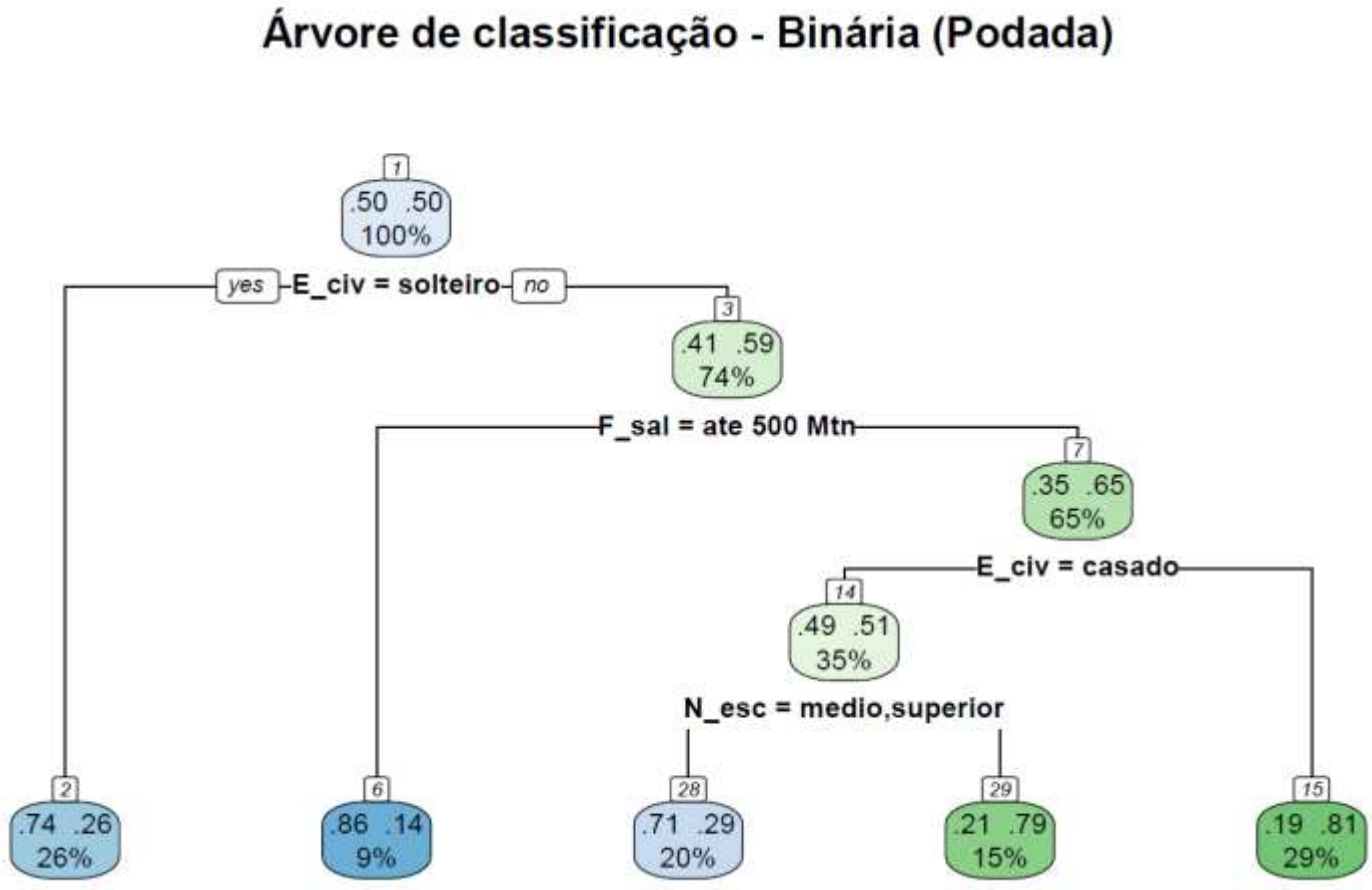
Ilustração 13: Gráfico de complexidade da árvore



Fonte: Autor

Observando o gráfico, tendo em conta a linha orizontal, percebe-se que depois de cinco folhas as subárvores não têm diferença significativa, sendo assim, vamos podar a árvore de modo a ter cinco folhas.

Ilustração 14: Árvore de classificação binária (podada)



Fonte: Autor

Embora esta árvore seja mais parcimoniosa que a ávore maximal (a primeira), dado que a sua obtensão resulta de um procedimento quase manual, que tomou em consideração apenas o valor de *cp,* surge a necessidade de questionar se não estariam lesados outros critérios por exemplo a robustez da árvore quando sujeita a dados não usados na construção do modelo. Para responder a estas questões vamos buscar a árvore óptima recorrendo à validação cruzada.

Validação Cruzada

No presente parágrafo vamos efectuar a validaçao cruzada de modo a obter informações realistas sobre a capacidade preditiva e a estabilidade dos modelos.

Para esse feito vamos recorrer à biblioteca *caret* do pacote R® e indicaremos o método como *rpart* (method="rpart") para construir árvores de classificação.

Vamos executar a função *train* que permite construir árvores podadas com base no parâmetro *cp* (custo de complexidade) e escolher a árvore podada óptima com base nas informações de validação cruzada repetida, podendo usar os parâmetros AUC e índice de kappa.

```
set.seed(1712)
if (colnames(dataVCR)[6] != "Sup_preferido") {
  stop("Certifique-se de que 'Sup_preferido' é a variável na 6ª coluna de 'dataVCR'.")
}

if (!exists("modeloEntropia") || !("cptable" %in% names(modeloEntropia))) {
  stop("Certifique-se de que 'modeloEntropia' está definido corretamente e contém 'cptable'.")
}

# Definir os parâmetros de treino
vcr <- train(
  y = dataVCR$Sup_preferido,
  x = dataVCR[,-6],  # Exclui a 6ª coluna (variável resposta)
  method = "rpart",
  control = rpart.control(minbucket = ceiling(0.01 * nrow(dataVCR))),
  metric = "AUC",
  parms = list(split = 'information'),
  tuneGrid = expand.grid(cp = modeloEntropia$cptable[,1]),
  trControl = trainControl(
    method = "repeatedcv",
    number = 5,
    repeats = 20,
    summaryFunction = multiClassSummary,
    classProbs = TRUE,
    returnResamp = "all",
    savePredictions = TRUE
  )
)
```

Para escolher a melhor árvore resultante da validação cruzada vamos recorrer aos comandos

results <- vcr$results, best_row <- which.max(results$AUC) e best_cp <- results$cp[best_row], que nos permitem determinar os resultados da validação cruzada, encontrar o índice da linha com o melhor desempenho (maior AUC) e Obter o valor de *cp* correspondente ao melhor desempenho, respectivamente, cujos resultados são:

Ilustração 15: Tabela 23: tabela de saidas de validação cruzada

```
> print(vcr)
CART

310 samples
 10 predictor
  2 classes: 'Game', 'Shoprite'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 20 times)
Summary of sample sizes: 248, 248, 248, 248, 248, 248, ...
Resampling results across tuning parameters:

  cp           logLoss    AUC        prAUC      Accuracy   Kappa       F1         Sensitivity  Specificity
  0.000000000  1.4642147  0.8094355  0.6047640  0.7377419  0.47548387  0.7407450  0.7522581    0.7232258
  0.006451613  1.2401410  0.8053122  0.5887119  0.7429032  0.48580645  0.7494312  0.7725806    0.7132258
  0.022580645  0.7788517  0.7805255  0.5807329  0.7569355  0.51387097  0.7558363  0.7554839    0.7583871
  0.083870968  0.6446345  0.6733143  0.3918971  0.6664516  0.33290323  0.6459731  0.6496774    0.6832258
  0.129032258  0.6721592  0.6072060  0.2677435  0.6067742  0.21354839  0.5814212  0.6258065    0.5877419
  0.251612903  0.6885759  0.5443548  0.1224263  0.5443548  0.08870968  0.6420695  0.8487097    0.2400000
  Pos_Pred_Value  Neg_Pred_Value  Precision  Recall     Detection_Rate  Balanced_Accuracy
  0.7350700       0.7487011       0.7350700  0.7522581  0.3761290       0.7377419
  0.7330159       0.7630218       0.7330159  0.7725806  0.3862903       0.7429032
  0.7630975       0.7603736       0.7630975  0.7554839  0.3777419       0.7569355
  0.6803347       0.6828239       0.6803347  0.6496774  0.3248387       0.6664516
  0.6223858       0.6532225       0.6223858  0.6258065  0.3129032       0.6067742
  0.5401647       0.6406437       0.5401647  0.8487097  0.4243548       0.5443548

AUC was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.
```

```
> best_row
[1] 1
> best_cp
[1] 0.006451613
```

Fonte: Autor

Analisando as saídas anteriores podemos constatar que, segundo os resultados da validação cruzada, a melhor árvore é aquela que corresponde à linha 1 com o cp=0,006451613, *AUC*=0,8053122, o que indica que o modelo tem uma boa capacidade de discriminação, ou seja, é capaz de distinguir efetivamente entre as duas categorias.

acurácia=0,7429032, o que significa que o modelo classifica corretamente 74,3% do total dos dados;

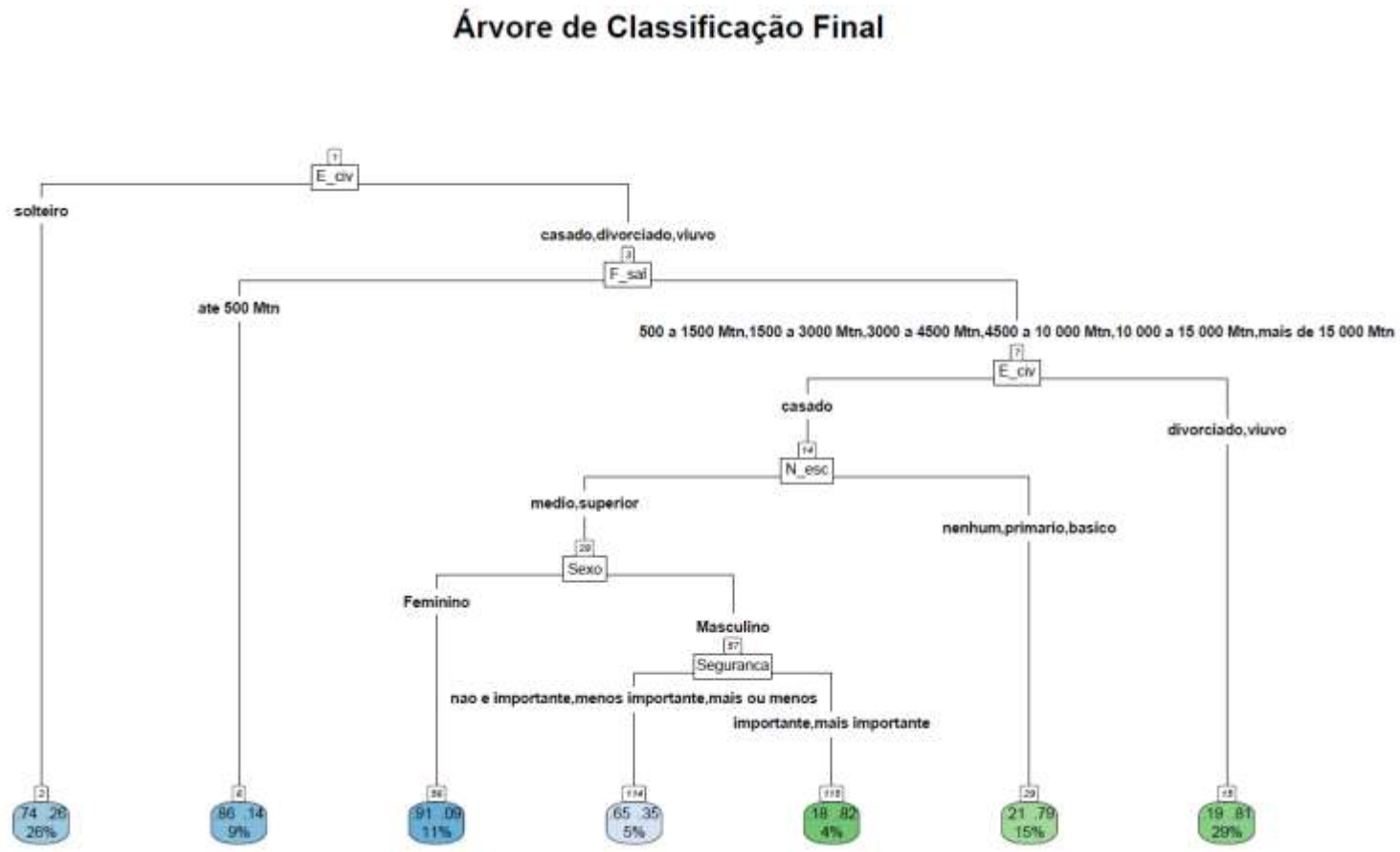
Kappa=0,48580645, este valor do coeficiente Kappa indica uma concordância moderada entre as classificações do modelo e as observações;

sensitividade=0,7725806, esta sensibilidade sugere que o modelo é capaz de identificar corretamente cerca de 77,3% dos casos positivos, o que é uma boa taxa de verdadeiros posivos;

especificidade=0,7132258, este valor de especificidade sugere que o modelo é capaz de identificar corretamente cerca de 82,1% dos casos negativos, que também representa um bome score.

E estas estatísticas descrevem o melhor modelo possível resultante do processo de validação crizada, cuja representação em forma da árvore vêde a seguir:

Ilustração 16: Figura 4: árvore de classificação binária (final)



Fonte: Autor

Interpretação da árvore:

Tomando em consideração que a nossa categoria de referêrncia nesta variável (supermercado preferido) Shoprite, tomas a seguinte interpretação para as folhas da árvore de classificação:

a) para folha 2 temos que, os solteiros são dos que menos procuram pelos supermercados Game pois, do universo destes somente 26% é que procuram por estes estabelecimentos;
b) para folha 6 temos que, apenas 14% das pessoas que não são solteiras e ganham até 500 Mtn, são favoráveis aos supermercados Game;
c) para folha 56 temos que, as mulheres casadas com nível de escolaridade médio ou superior que ganham mais de 500 Mtn, 9% destes são os favoráveis ao supermercado Game, sendo este grupo oque menos é favorável aos supermercados Game;

d) para folha 114 temos que, 35% dos homens casados, com nível de escolaridade que varia de medio a superior, que ganham mais de 500 Mtn eque não levam em consideração alguma a segurança, são favoráveis ao supermercado Game;
e) para folha 115 temos que, 82% dos homens referidos na alínea anterior com a particularidade destes acharem que a segurança é importante ou mais importante, são favaráveis ao supermercado Game;
f) para folha 29 temos que, 79% das pessoas casadas com o nível de escolaridade que vai até básico, ganhando acima de 500 Mtn, são favoráveis ao supermercado Game;
g) Para folha 15 temos que, 81% das pessoas que são viúvas ou divorciadas, que ganham mais de 500 Mtn, são favoráveis aos supermercados Game.

Observe que embora esta árvore seja menos parcimoniosa que a anterior, ela é maisparcimoniosa que a maximal (possuindo 7 folhas contra 12), e resulta de processo de construcão fiável (validação cruzada). Sendo assim, assumimo-la como óptima e vale para o efeito de previsão e interpretação da base de dados.

Conclusão:

Tendo em conta que a nossa categoria de referência para a variável de decisão (supermercado preferido) é Shoprite, com as interpretações acima, podemos concluir que dos inqueridos, os menos favoráveis ao supermercado Game são as mulheres casadas com nível de escolaridade médio ou superior, e que ganham mais de 500 Mtn ou, pessoas de qualquer sexo que ganhem até 500 Mtn. De outro lado, uma parte considerável dos favoráveis ao supermercado game é constituida por pessoas que não são solteiras e ganham mais de 500 Mtn, com a maior destaque para homens com níveis de escolaridade acima do básicoe que valoriza as consições de segurança. Estas conclusões vêm reforçar os resultados do modelo de *regressão logística binária* em que as chances de um viúvo, casado, divorciado escolher supermercado game eram muito superiores a 100% em comparação com solteiros(categoria de referência). Da mesma forma que, quanto maior for a faixa salarial, nível de escolaridade, as chances proporcionais têm a mesma tendencia de ser um múltiplo superior a unidade em comparação com as respectivas categorias de referência.

Árvore de Classificação para modelo de Classificação Ordial

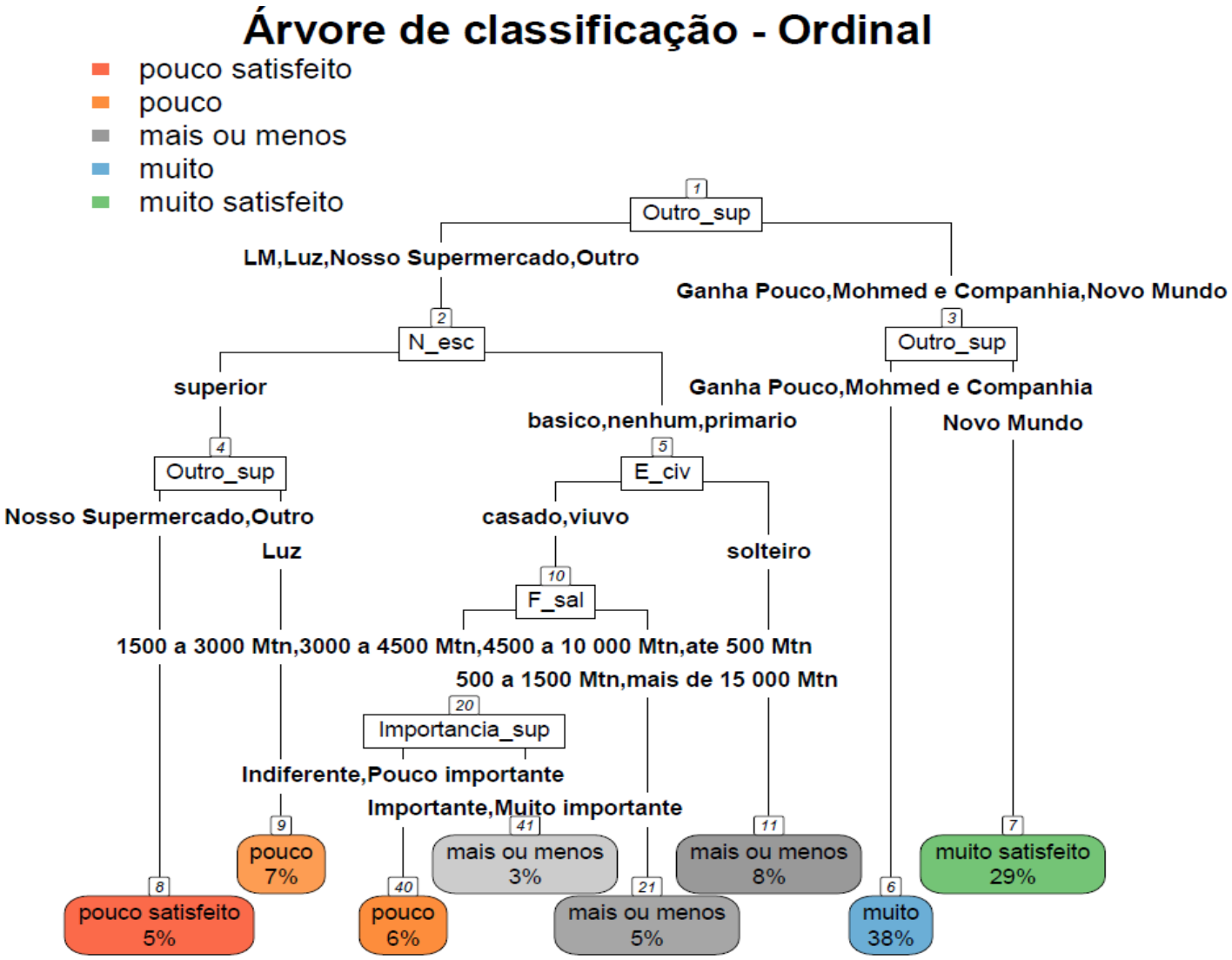
No presente parágrafo queremos criar um modelo que permite classificar os inqueridos em diferentes grupos, conforme os níveis de satisfação alcançados, a saber: *pouco satisfeitos, pouco, mais ou menos,*

muito e muito satisfeitos, em alternativa ao modelo de *Regressão logística ordinal*. Para esse feito vamos usar a funão *rpart* do R®. À semelhança do modelo de *regressão logística ordinal*, vamos trabalhar com as variáveis : carecterísticas do indivíduo (sexo, faixa etária, estado civíl, faixa salarial, e estado civíl), ordem de valores (Empatia, Segurança, Presteza, Confiabilidade e Tangíveis), Informação geral sobre o serviço do supermercado e a própria variável resposta que é o nível de satisfação.
Os dados são repartidos em duas partes, uma para treinamento (80%) e outra para prova ou teste (20%).

a) Criação e interpretação da árvore

Para criação da árvore de classificação, em conformidade com a literatura usada nesta pesquisa, vamos aqui tambem usar o critério de *Entropia.* Com o recurso à função *rpart* do pacote do mesmo nome do software R®, construimos a árvore de classificação a seguir:

Ilustração 17: Árvore de classificação ordinal (maximal)



Fonte: Autor

Construida a árvore de classificação ordinal pelo critério de entropia, podemos ver que ela possue 10 folhas e tem de comprimento 5, de seguida iremos estudar a eficiência da mesma em termos de capacidade preditiva e robustez. Executando os comandos R® para o cálculo da matriz de classificação,

probs <- predict(modeloEntropiaOrd, data_test, type = "class") *# Realiza previsões no conjunto de teste*
cm <- confusionMatrix(data = as.factor(probs), reference = data_test$Nivel_satisfacao) *# Cria matriz de confusão*

temos:

Tabela 23: Matriz confusão/classificação da árvore de classificação ordinal

```
> cm$table
                 Reference
Prediction        pouco satisfeito pouco mais ou menos muito muito satisfeito
  pouco satisfeito                5     1             0     0                0
  pouco                           0     8             0     0                0
  mais ou menos                   0     1             9     0                1
  muito                           0     0             0    27                2
  muito satisfeito                0     0             0     0               22
```

Fonte: Autor

Analizando a presente matriz de classificação ou confusão, nota-se poucas ocorrências dos erros do tipo I e II, ito é há poucos dados clasificados para uma categoria quando pertencem a outra, e caso em que isso ocorre entende-se como quase evento uma vez que esse erroocorre em categorias consecutivas com uma excepção de uma observação que foi classificada na categoria *mais ou menos* quando na verdade ela pertence a *muito satisfeito.* Para melhor discrição da eficiência do modelo vejamos abaixos as estatísticos: acurácia, eficiência e especificidade.

```
Overall Statistics

               Accuracy : 0.9342
                 95% CI : (0.8531, 0.9783)
    No Information Rate : 0.3553
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9103
```

Como se pode ver, de uma forma geral, contamos com uma acurácia de aproximadamente 93%, istá é, a proporção de eventos correctamente classificados, e um índice deKappa igual a 0,9301 que pode ser classificado como de uma concordância quase perfeita.

Na tabela abaixo analizemos a robustez de modelo tendo em conta a sensitividade e especificida por cada categoria.

Tabela 24: Análise da robustez da árvore de classificação ordinal

Statistics by Class:

	Class: pouco satisfeito	Class: pouco	Class: mais ou menos	Class: muito
Sensitivity	1.00000	0.8000	1.0000	1.0000
Specificity	0.98592	1.0000	0.9701	0.9592
Pos Pred Value	0.83333	1.0000	0.8182	0.9310
Neg Pred Value	1.00000	0.9706	1.0000	1.0000
Prevalence	0.06579	0.1316	0.1184	0.3553
Detection Rate	0.06579	0.1053	0.1184	0.3553
Detection Prevalence	0.07895	0.1053	0.1447	0.3816
Balanced Accuracy	0.99296	0.9000	0.9851	0.9796
	Class: muito satisfeito			
Sensitivity	0.8800			
Specificity	1.0000			
Pos Pred Value	1.0000			
Neg Pred Value	0.9444			
Prevalence	0.3289			
Detection Rate	0.2895			
Detection Prevalence	0.2895			
Balanced Accuracy	0.9400			

Fonte: Autor

Sensitividade

Para as categorias: *pouco satisfeito, mais ou menos e muito,* o modelo tem a capacidade de detectar a observações que realmente são eventos a 100%, 80% para *pouco* e 88% para *muito satisfeito.*

Especificidade

Para as categorias: *pouco* e *muito satisfeito* o modelo tem a capacidade de detectar a observações que não são eventos de 100%, 99% para *pouco satisfeito,* 97 para *mais ou menos* e 96% para *muito.*

Com estes valores de estatísticos podemos afirmar que o modelo tem um excelente desempenho e é bem robusto.

Poda da árvore

De modo a reduzir o super ajuste do modelo (árvore), vamos execur a poda da árvore construida anteriormente (maximal), para obter uma árvore com menos folhas (mais parcimoniosa), se esta for a melhor alternativa.

Primeiramente vamos construir a tabela de *custo de complexidade - cp* e o respectivo gafico que nos permitem visualizar as diferentes subarvores e os respectivos erros associados, para esse feito vamos executar os comandos R® abaixo.

modeloEntropiaOrd$cptable
plotcp(modeloEntropiaOrd)

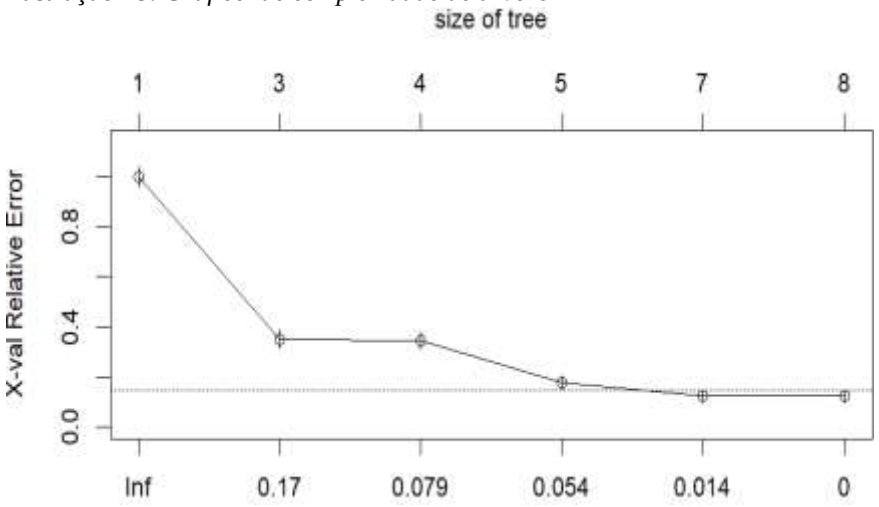
Tabela 25: Tabela de complexidade de árvore classificação ordinal

	CP	nsplit	rel error	xerror	xstd
1	0.326732673	0	1.0000000	1.0000000	0.04152937
2	0.084158416	2	0.3465347	0.3514851	0.03662654
3	0.074257426	3	0.2623762	0.3465347	0.03644370
4	0.039603960	4	0.1881188	0.1782178	0.02792507
5	0.004950495	6	0.1089109	0.1237624	0.02373341
6	0.000000000	7	0.1039604	0.1237624	0.02373341

Fonte: Autor

Com esta tabela de complexidade podemos escolhher a subárvore que minimiza o erro e possivelmente o overfitting, mas uma melhor visualização podemos obter a partir do gráfico de complexidade a seguir.

Ilustração 18: Gráfico de complexidade de árvore



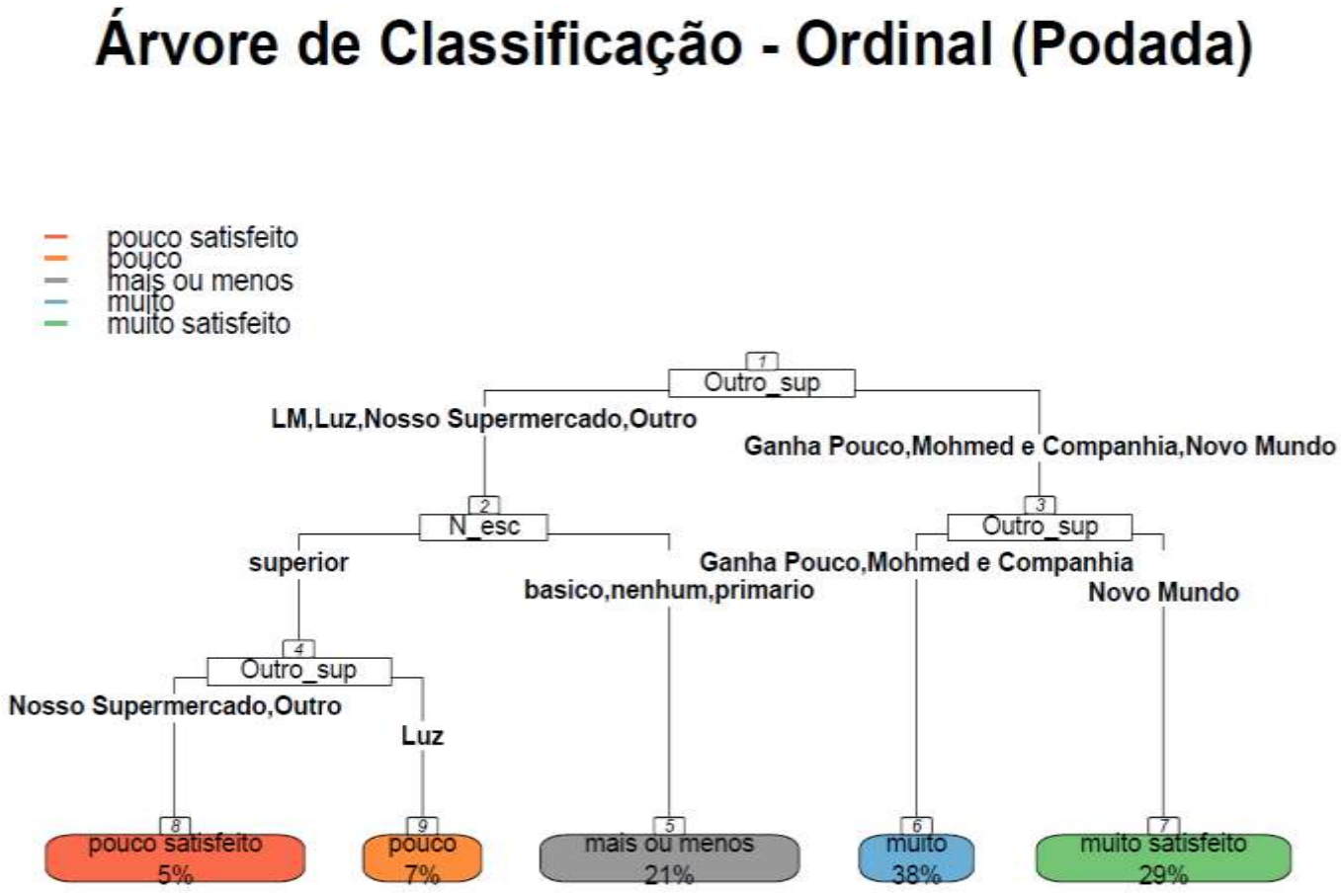
Fonte: Autor

Observando o gráfico, tendo em conta a linha orizontal, percebe-se que depois de cinco folhas as subárvores não têm diferença significativa, sendo assim, vamos podar a árvore de modo a ter cico folhas.

Executando o comando R® para construção da árvore com indicação com indicação do número de folhas, temos:

modeloEntropiaOrdPodado <- prune(modeloEntropiaOrd, cp = modeloEntropiaOrd$cptable[4,1])
rpart.plot(modeloEntropiaOrdPodado,main = "Árvore de Classificação - Ordinal (Podada)",
type=5,extra = 100,nn=TRUE,tweak=1.0)

Ilustração 19: Árvore de classificação ordinal (poda)



Fonte: Autor

Depois da poda temos uma árvore com cinco folhas e de comprimento três.

Validação Cruzada

No presente parágrafo, à semelhança do que ocorreu no modelo de *regressão logística binária*, vamos efectuar a validaçao cruzada de modo a obter informações realistas sobre a capacidade preditiva e a estabilidade dos modelos. com recurso à biblioteca *caret* do pacote R® e indicaremos o método como *rpart* (method="rpart") para construir árvores de classificação.

Vamos executar a função *train* que permite construir árvores podadas com base no parâmetro *cp* (custo de complexidade) e escolher a árvore podada óptima com base nas informações de validação cruzada repetida e, para tal vamos usar os parâmetros AUC.

```
#Validacao Cruzada
dataVCR<-data_train
dataVCR$Nivel_satisfacao<-factor(dataVCR$Nivel_satisfacao, levels=rev(levels(dataVCR$Nivel_satisfacao)),
                  labels=c("pouco satisfeito", "pouco", "mais ou menos", "muito", "muito satisfeito"))
table(dataVCR$Nivel_satisfacao)/nrow(dataVCR)
levels(dataVCR$Nivel_satisfacao)
levels(dataVCR$Nivel_satisfacao) <- make.names(levels(dataVCR$Nivel_satisfacao), unique = TRUE)
levels(dataVCR$Nivel_satisfacao)

set.seed(1712)
vcr<-train(y=dataVCR$Nivel_satisfacao, x=dataVCR[,-11],
           method = "rpart",
           control = rpart.control(minbucket=ceiling(0.01*nrow(dataVCR))),
           metric="AUC", parms=list(split='information'),
           tuneGrid = expand.grid(cp=modeloEntropiaOrd$cptable[,1]),
           trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                    summaryFunction=multiClassSummary,
                                    classProbs=TRUE, returnResamp = "all",
                                    savePredictions = TRUE))
```

Para escolher a melhor árvore resultante da validação cruzada vamos recorrer aos comandos

results <- vcr$results;

best_row <- which.max(results$AUC) e

best_cp <- results$cp[best_row],

que nos permitem determinar os resultados da validação cruzada, encontrar o índice da linha com o melhor desempenho (maior AUC) e Obter o valor de *cp* correspondente ao melhor desempenho, respectivamente, cujos resultados são:

Tabela 26: Resultado da validação cruzada

```
> results
          cp   logLoss       AUC      prAUC  Accuracy     Kappa   Mean_F1 Mean_Sensitivity Mean_Specificity
1 0.000000000 1.1511635 0.9642121 0.20816032 0.9166158 0.8866584 0.8842465        0.8832224        0.9789666
2 0.004950495 1.0928214 0.9634664 0.19759231 0.9187127 0.8894946 0.8859722        0.8848571        0.9795430
3 0.039603960 0.8402441 0.9506314 0.12676635 0.8843599 0.8427860 0.8356809        0.8384532        0.9712012
4 0.074257426 0.7760963 0.9366228 0.08052951 0.8454434 0.7891958 0.7986428        0.7364373        0.9620624
5 0.084158416 0.7971976 0.9182986 0.06448269 0.8139180 0.7445395 0.7625745        0.6532151        0.9536516
6 0.326732673 1.1336990 0.7212712 0.01057535 0.5901044 0.3867544       NaN        0.3753000        0.8795232
```

```
> best_row
[1] 1
> best_cp
[1] 0
```

Fonte: Autor

Analisando as saídas anteriores podemos constatar que a melhor árvore é aquela que corresponde aproximadamente aos estatísticos:

cp=0, que entendemos como um valor correspondente a uma arvore menos complexa;

AUC=0,9642, o que indica que o modelo tem uma boa capacidade de discriminação, ou seja, é capaz de distinguir efetivamente categorias diferentes.

acurácia=0,9166, o que significa que o modelo classifica corretamente 91% do total dos dados; *Kappa=0,8867,* este valor do coeficiente Kappa indica uma concordância forte entre as classificações do modelo e as classificações reais;

sensitividade (média)=0,8832, esta sensibilidade sugere que o modelo é capaz de identificar corretamente cerca de 88,32% dos casos positivos, o que é uma boa taxa de verdadeiros posivos;

especificidade (média)=0,9789, este valor de especificidade sugere que o modelo é capaz de identificar corretamente cerca de 97,89% dos casos negativos, que também representa um bom score.

Importância das Variáveis

De seguida, vamos analizar a importância das variáveis de modo a aferir quais são as variáveis que dão mais informações na criação do modelo (árvore final). Para tal vamos executar os comandos de R®, *modeloEntropiaOrdFinal$variable.importance* e *percentagens <- dados / sum(dados) * 100*que dão o resuldo abaixo:

Outro_sup	N_esc	E_civ	F_sal	Importancia_sup
81.515350	10.367645	3.756590	2.710525	1.649890

Portanto, a variável mais importante é outro supermercado (*outro_sup)* 81,52%, seguida pela variável nível de escolaridade (*N_esc*) com 10,37%, estado civíl (*E*_civ) 3,76, faixa salarial com 2,71% e importância de supermercado (*importancia_sup) 1,65*%.

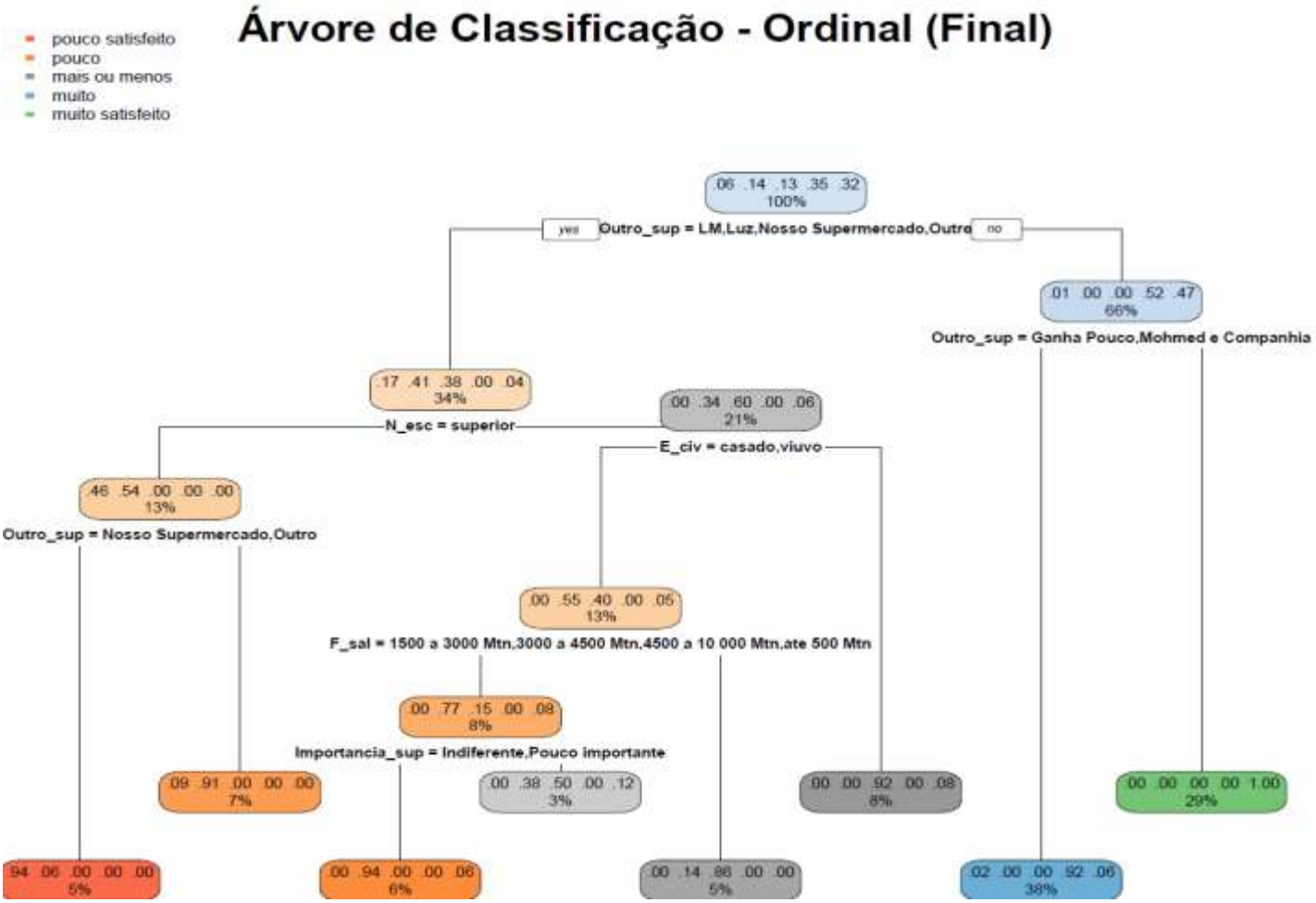
De seguida, para uma melhor visualização, temos a representação gráfica da importância das variáveis.

Ilustração 20: Gráfico de importância de variáveis

Fonte: Autor

E com base nass estatísticas de saída da validação cruzada, que descrevem o melhor modelo possível , a representação em forma de árvore é:

Ilustração 21: Árvore de classificação ordinal (final)



Fonte: Autor

Observando a árvore de classifição tida como óptima através da validação cruzada, podemos constatar que esta tem mais folhas que aquela que resultou de uma poda manual. E como os parâmetro que ditaram a sua construção, pode se observar aida que a mesma é a que foi construida inicialmente. Portanto são transferidas todas as avaliações do desempenho, estabilidade e robostez daquela.

4.5 Interpretação da Árvore de Classificação Ordinal Final

- Tendo como preferência de outro supermercado (para alem de Game ou shoprite), os supermercados Nosso supermercado e Outro supermercado e, ter nível de escolaridade superior, tem 94% de probabilidade de sair *pouco satisfeito* com os serviços prestados nos superercados *Game ou Shorite,* o que significa que tem 6% de chances de alcançar a categorias acima desta (*pouco*).
- Tendo como preferência de outros supermercados os supermercados Luz e LM e, ter nível de escolaridade superior, tem 91% de probabilidade de sair com nível de satisfação *pouco* com os serviços prestados nos superercados *Game ou Shorite, 9%* de chances de alcançar categoria abaixo daste.
- Tendo como preferência de outro supermercado, os supermercados Luz, LM, Nosso supermercado e Outro supermercado e, ter nível de escolaridade não superior, sendo casado ou viúvo, ganhando salário diferentes de 500 a 1500 Mtn ou acima de 10000 Mtn e considerando a importância do supermercado indiferenete ou pouco importante,te m 94% de probabilidade de sair com o nível de satisfação até *pouco* com os serviços prestados nos superercados *Game ou Shorite,* e 6% de chances de alcançar a categoria *muito satisfeito*.
- Tendo supermercado Luz, LM, Nosso Supermercado, Outro supermercado, nível de escolaridade que varie de nenhúm a básico e solteriro ou divorciado, tem 92% de probabilidades de alcançar um nível de satisfação *mais ou menos,* e porque o modelo acumula probabilidades, tambem podemos concluir que este tipo de pessoas têm 8% de chances de alcansar níveis de satisfação até *muito satisfeito.*
- Tendo supermercado Ganha pouco e Mohmed e Companhia como outra preferência, tem 92% de probabilidades de sair com nível de satisfação até *muito* e 6% de probabilidades de alcançar até *muito satisfeito* como nível de atisfação com os supermercados *Game ou Shoprite*.
- Tendo supermercado Novo Mundo como outra preferência, tem 100% de probabilidades de sair com nível de satisfação *muito satisfeito* com os supermercados *Game ou Shoprite*.

4.6 Escalonamento Multidimensional (MDS)

Com o objectivo de estudar as relações existentes entre aquelas variáveis que não foram tidas como relevantes na construção dos modelos anteriores (RLB & RLO), isto é, as relações existentes entre as

persepções dos diferentes clientes dos supermercados da província e cidade de Maputo, nos seus diferentes aspectos em termos de qualidade dos serviços fornecidos, nesta secção vamos aplicar a técnica de Escalonamonto Multidiensional (MDS) e as respectivas avaliações e interpretações de modo a obter as respostas das questões acima espostas.

Em jeito de extensão ao escalonamento multidimensional tradicional, que busca estudadar as similaridades entre as observações,como nos refirimos na metodologia, no nosso trabalho vamos nos focar na busca de similaridades entre as variáveis.

Trabalhando com a nossa base de dados "*satisfação com o supermercado Game ou Shoprite*", com o recurso ao sotwear R®, usando as livrarias: cluster, StatMatch, psych, smacof, OneR, DescTools, Questionr e ClustOfVar.

Estando diante de variáveis categóricas ordinais, para o cálculo de MDS vamos usar a matriz de correlação de variáveis, usando o coeficiente de correlação τ_β de Kendall, segundo o artigo "*Measures of Association How to Choose?* " de Harry Khamis, Phd.

Usando os comandos R® abaixo, vamos determinar a matriz de correlação de Kendall, a correspondente matriz de dissimilaridade e o MDS.

```
matrizCor<- cor(dados, method = "kendall") #Kendal tau-beta
matrizDisim <- sim2diss(matrizCor, method = 1)
MDSvariables_V<- mds(matrizDisim, type = "ordinal")
MDSvariables_V
```

Cujos resultados são:

```
Call:
mds(delta = matrizDisim, type = "ordinal")

Model: Symmetric SMACOF
Number of objects: 17
Stress-1 value: 0.247
Number of iterations: 78
```

Em resumo, a análise de MDS utilizou 17 objetos (variáveis), alcançou um valor de Stress de 0,247 e convergiu após 78 iterações. Todavia, observando o valor de stress, tendo em conta a regra de ouro, diríamos que a qualidade da representação gráfica é insuficiente para tirar conclusões robustas. Neste

caso, vamos recorrer ao teste das permutaçãoes para ver se, uma vez que a dimensão de matriz de dissimilaridades influencia no valor de *stress*, e outras características da mesma, ão seria o caso da ineficiência desta regra.

E o código para esse este teste em R® é:

```
set.seed(123)
permCol <- permtest(MDSvariables_V, nrep = 500, verbose = FALSE)
permCol
```

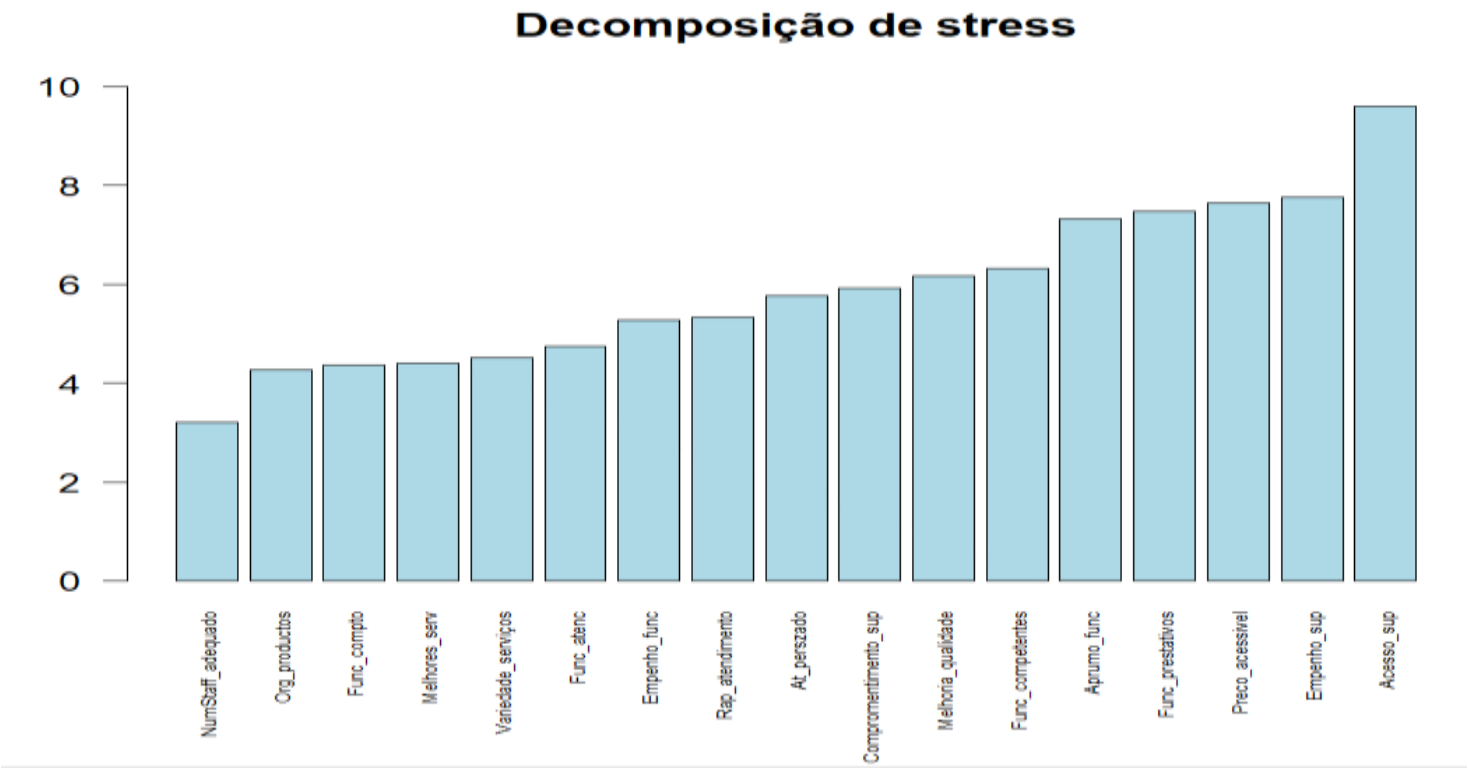
cujo resultado é:

```
Call: permtest.smacof(object = MDSvariables_V, nrep = 500, verbose = FALSE)

SMACOF Permutation Test
Number of objects: 17
Number of replications (permutations): 500

Observed stress value: 0.247
p-value: 0.004
```

e porque o *p-valor* é 0,004, inferior a 0,05, podemos concluir que o stress obtido é suficientemente pequeno e dele podemos tirar todas conclusões necessárias.

De seguida vamos determinar a contribuição de cada variável para o stress, de modo verificar quais são as variáveis que têm melhor ou pior representação.

Ilustração 22: Gráfico de decomposição de stress



Fonte: Autor

A variável com pior representação é o acesso ao supermercado (Acess_sup) com pouco mais de 10% do stress, tendo como consequência deste facto, falsas inferências sobre a natureza das relações entre esta variávelou outras com representação similar e as demais, impactando qualquer análise subsequente ou decisão tomada relativamente a esta variável. Enquanto que o número de staff (NumStaff_adeq) com pouco menos de 4%, é o que apresenta melhor representação, o que garante que a estrutura e as relações nos dados originais que envolve esta ou outras variáveis com representação similar sejam mantidas, levando a análises mais precisas e interpretações confiáveis, o que favorece uma tomada de decisões eficaz.

Ilustração 23: Gráfico de configuração de pontos (variáveis)

Fonte: Autor

Observsando o grafico acima, pode se verificar que os supermercados que apresentam serviços Variados, tambem apresenta funcionários empenhados; os que apresentam funcionárias mais prestitivos tambem se notabilizam-se em termos de atendimento rápido; os que apresentam funcionários competentes tambem têm bom desempenho; o que nos permite concluir que as variáveis que representam estes atributos são relacionadas, embora não tenhamos elementos para concluir se esta relação é directa ou inversa. Ao contrário, podemos verificar através do distanciamento das variáveis que, supermercados que têm preços acessíveis, não se notabilizam por ter boa organização dos productos; aos que apresentam serviços variados não se natabilizam por atendimento personalizado.

Cluster de variáveis

No presente parágrafo vamos criar agrupamentos de variáveis que temos estado a analizar a sua similaridade/dissimularidade. Para esse feito vamos utilizar as saidas de MDS, de modo a aferir até que

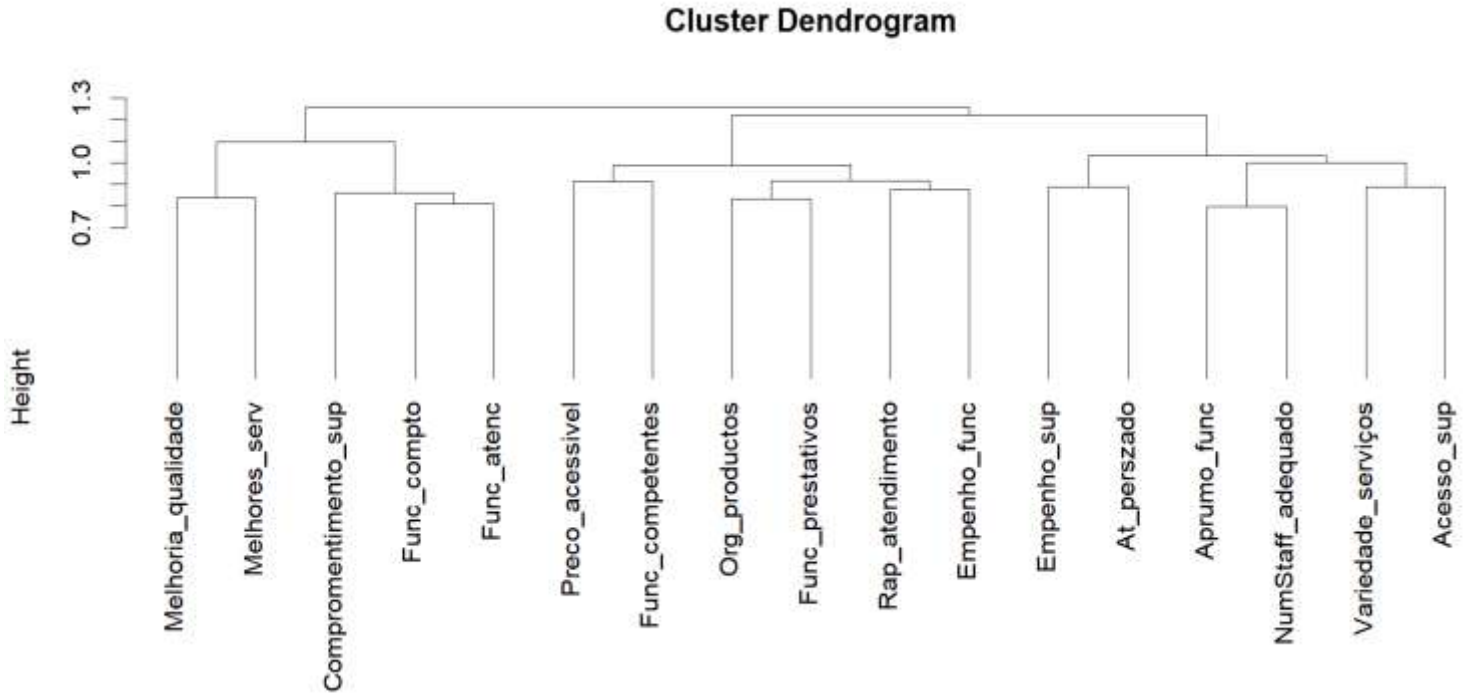
ponto as variáveis partilham ou não informaçãoes. Para tal, utilizaremos a biblioteca ClustOfVar do pacote do mesmo nome, para criar um dendrograma que representa os agrupamentos.

Executando os comandos em R®,

```
library(ClustOfVar)
tree <- hclustvar(X.quanti = dados, X.quali =NULL)
plot(tree)
```

temos:

Ilustração 24: Dendrograma de Clusters de variáveis



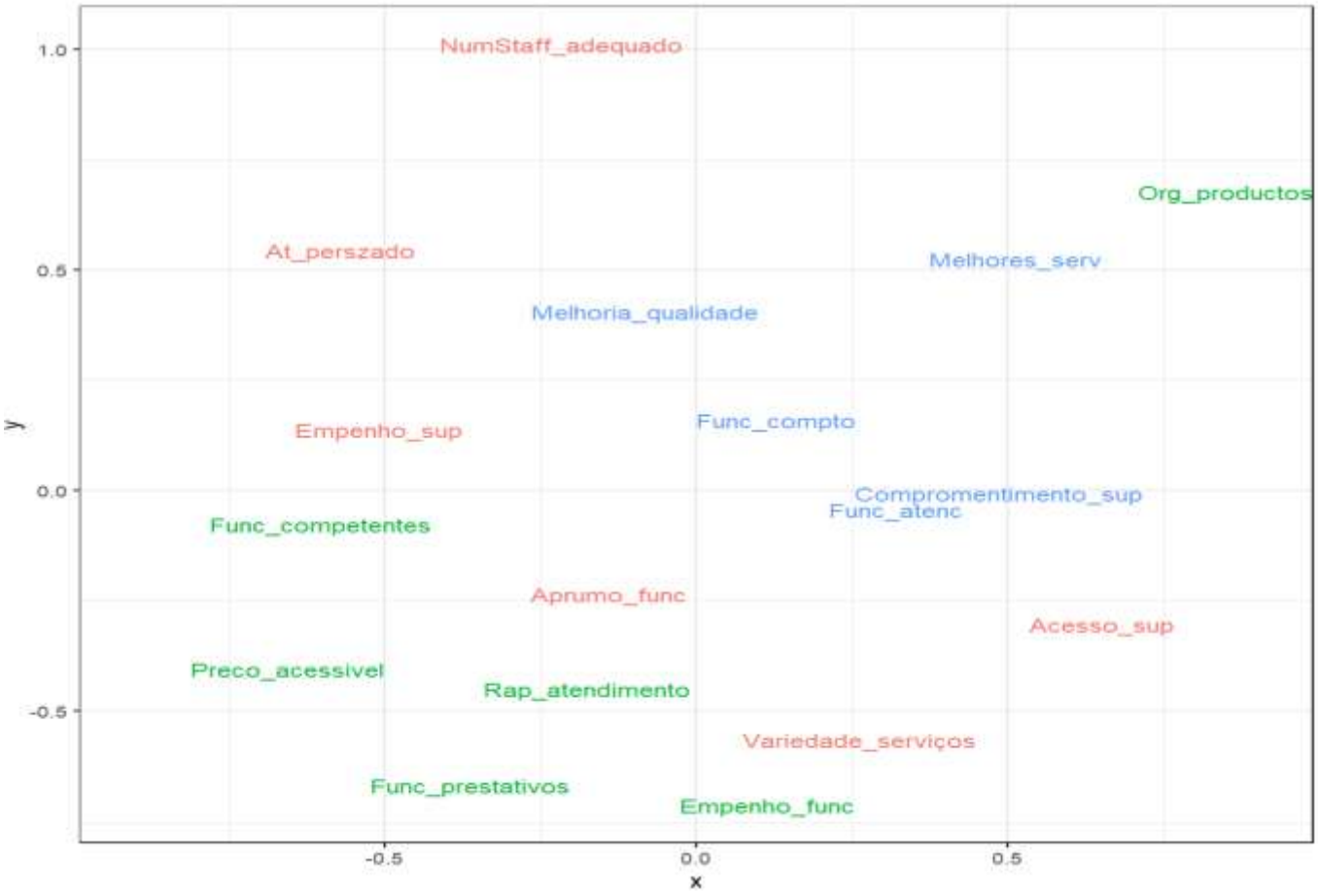
Fonte: Autor

Conforme o dendrograma nos mostra, temos os agrupamentos:

a) **Qualidade de serviços -** Melhoria_qualidade, Melhoria_serv, Comportamento_sup, Func_comp e Func_atenc;
b) **Organização de Supermercado** -Preco_acecivel, Func_competentes Org_productos, Func_Prestativos, Rap_atendimento e Empenho_func;
c) **Qualidade do Staff -** Empenho_sup, At_perszado, Aprumo_func e NumStff_adequado, Variedade_serviços e Acesso_sup.

De seguida vamos apresentar o grafico de configuração usando cores para uma melhor leitura.

Ilustração 25: Gráfico de configuração de pontos (à cores)



Fonte: Autor

Observando os agrupamentos nas suas diferentes cores percebe-se que elementos do mesmo agrupamento (definidos pela mesma cor), encontram-se demasiadamente afastados. Para corrigir isso temos que definir um número adequado dos agrupamentos, para tal, vamos utilizar o estudo de estabilidade, o que é feito obtendo várias amostras de bootstrap do conjunto de dados, aplicando o processo de cluster hierárquico e calculando o índice Rand ajustado para cada um dos agrupamentos obtidos e depois representar os resultados em diagramas de caixas para efeito de análise de estabilidade. Para tal, executemos os comandos R®:

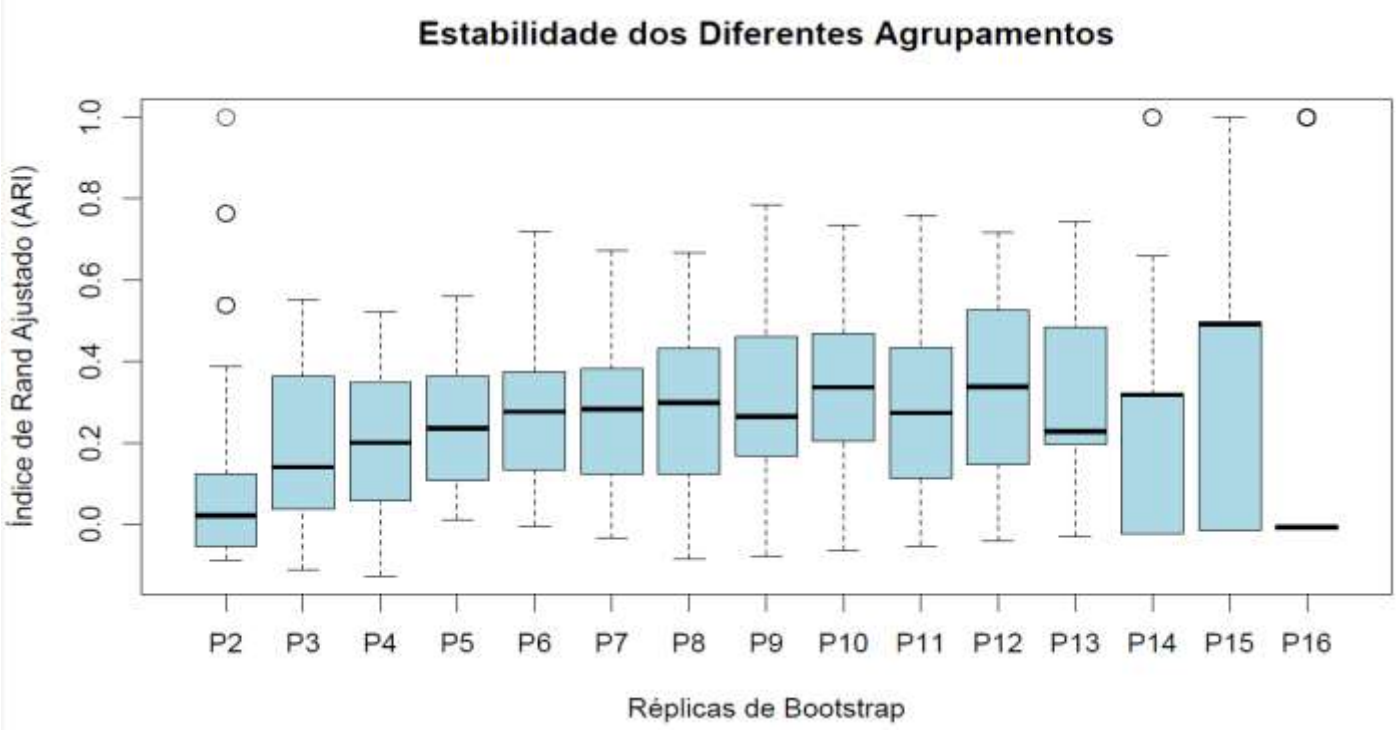
set.seed(12345)

stab <- stability(tree, B = 50, graph = F)

boxplot(stab$matCR, main = "Estabilidade dos diferentes agrupamentos")

e temos como resultado:

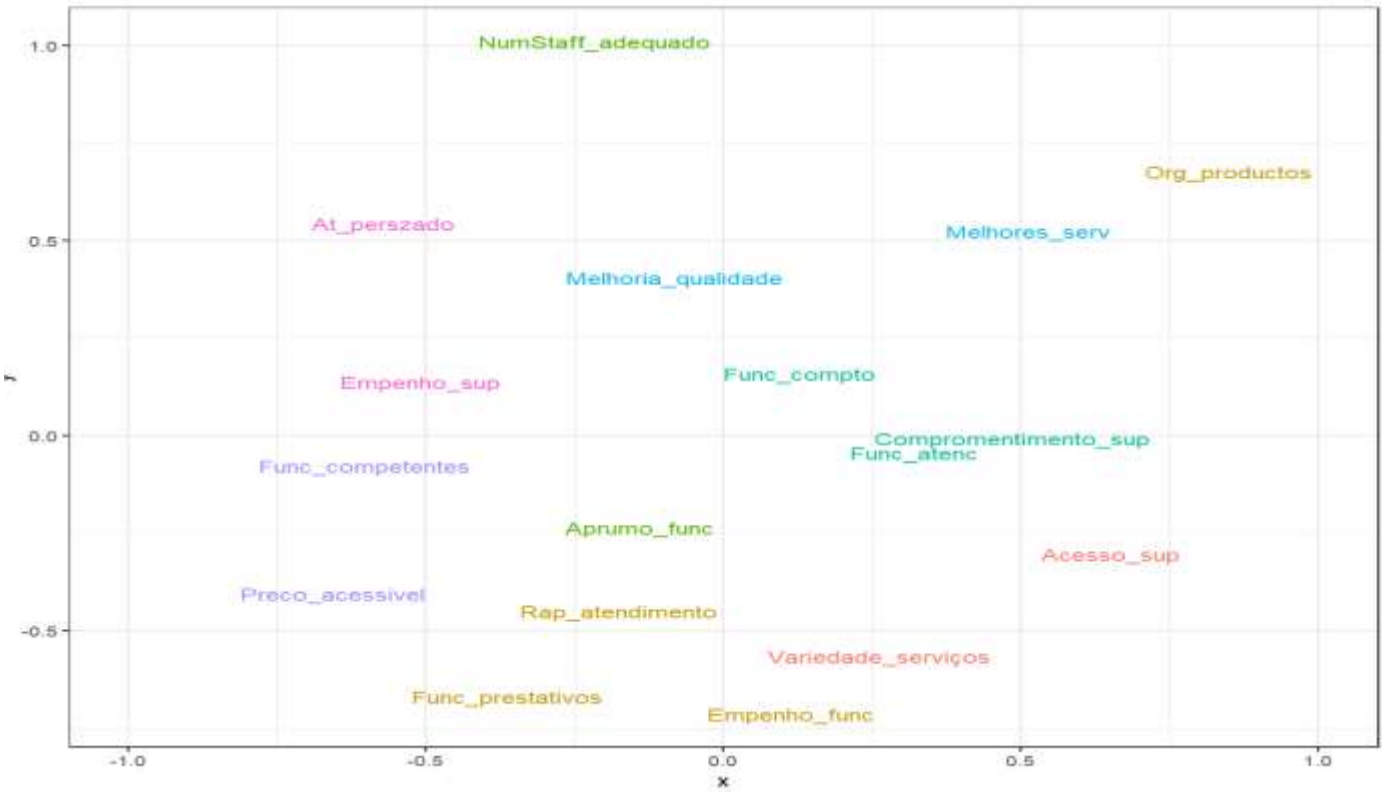
Ilustração 26: Gráfico de estabilidade dos agrupamentos



Fonte: Autors

Nota-se alguma estabilidade em oito agrupamentos, dado que estes apresentam a mediana do índice de rand Ajustado (ARI) relativamente elevado, isto é, maior que zero ($-1 \leq ARI \leq 1$), um baixo desvio interquartil, o que indica uma boa estabilidade e por fim a ausência de *outliers* nestas amostras indicam a ausência de diferenças significativas entre estes agrupamentos com o original, pelo que, construiremos de seguida um agrupamento com oito grupos. Executando de novo o comando *cutreevar*, mas com indicação para 8 agrupamento, temos:

Ilustração 27: Gráfico de configuração de pontos (à cores)

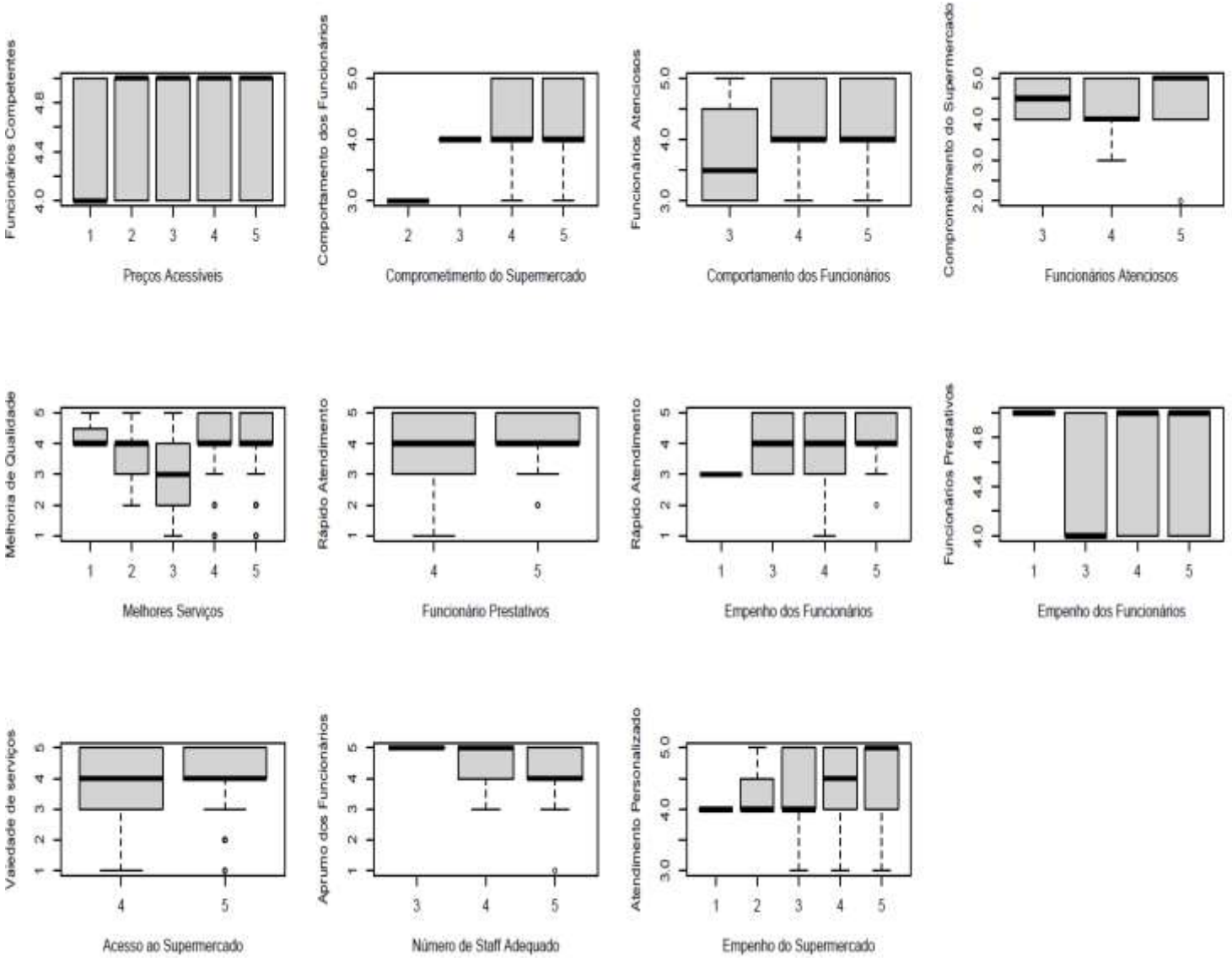


Fonte: Autor

Como se pode observar, os agrupamentos aqui obtidos concordam bastante com os obtidos no dendrograma, com excepção de dois agrupamentos que apresentam um elemento demasiadamente afastado.

Para perceber melhor a relação existente entre as variáveis agrupadas, vamos construir gráfico *boxlot* para auxiliar nesta análise.

Ilustração 28: Gráfico de relaçõesentre variáveis dos agrupamentos



Fonte: Autor

Observando os gráficos, percebe-se que apesar de algumas irregularidades nos seus padrões, nota-se que categorias altas de uma dada variável correspondem a categorias alta da outra variável com que se agrupa ou, mantêm constante as suas categorias quando a outra varia, tal é o caso de Func_competentes e Preco_acessivel, Rap_atendimento e Func_prestativos e, Variedade_servicos e Acesso_sup. Mas temos a excepção de NumStaff_adeq em função de Aprumo_func que apresentam uma tendência de

proporcionaliade inversa, o que indica que categorias altas de uma variável correspondem a categorias baixas da outra.

Pelo que, com estes resultados, entendemos que na óptica dos inqueridos, temos a seguinte leitura:

- os Supermercados que têm funcionários com categorias altas de competência, tambem são tidos como de categorias altas na prática de preços acessíveis;
- os supermercados que têm funcionários com categorias altas de empenho, estes têm atendimento mais rápido e são máis prestativos;
- Se o supermercado tem alta categoria de empenho, então possue atendimento personalizado tambem de gategoria alta;
- Alto nível de empenho dos funcionários está relacionado com alto nível de atenção dos funcionários e o comprometimento do supermercado;
- A variedade dos serviços disponíveis está relacionado à qualidade de acesso ao supermercado;
- E por fim , melhoria de qualidade está relacionado com melhoria de serviços.

5 Conclusãoes e Recomendações

5.1 Conclusões

Do estudo realizado com base nos modelos Regressão logística binária, Regressão logística ordinal, com as respectivas árvores de classificação, sobre a base de dados *Satisfação com o Supermercado Game ou Shoprite*, em resposta às perguntas de pesquisa, temos as seguintes conclusões:

a) São mais favoráveis aos Supermercado Game os clientes cujas características são:
   - os que ganham mais de 500 Mtn sendo, divorciados ou viúvos;
   - casados que não têm além do ensino básico;
   - e casados do sexo masculino com nível médio ou superior e que valorizam muito a questão de segurança.

Destes, em termos salariais destacam-se mais os que ganham entre 3000 e 10000 Mtn, em relação aos que ganham menos de 500 Mtn, os do sexo feminino em relação aos do sexo masculino, os viúvos em relação aos solteiros e os que têm ensino primário em relação aos que não têm nenhum nível de escolaridade. E, cruzando estas conclusões provenientes do modelo de regressão logística binária e árvore de classificação binária, podemos concluir em última análise que, as viúvas com nível de escolaridade primária, que ganham entre 3000 e 10000 Mtn, são as mais favoráveis ao supermercado Game em detrimento de Shoprite.

b) Já para o nível de satisfação dos clientes com os supermercados Game e Shoprite da cidade e províncias de Maputo, tendo em conta os resultados do modelo de regressão logistica ordinal e árvore de classificação ordinal, temos que:

- Indivíduos do sexo feminino, divorciados, com nível de escolaridade primária e os que ganham entre 4500 e 10000 Mtn, são os mais susceptiveis de alcançar elevados níveis de satisfação, tendo em conta as respectivas categorias de referência. A árvore de classificação ordinal apresenta as conclusões: são mais susceptíveis a alcançar elevados níveis de satisfação as pessoas que têm como outras preferencias os supermercados Novo Mundo (no topo), Ganha pouco e Mohmed e Companhia. Enquanto que os que têm preferência de outro supermercado, LM, Luz, Nosso supermercado e Outro supermercado, tem apenas até 12% de probabilidades de alcançar nível de satisfação *muito satisfeito.* Esta última parte permite-nos concluir que, estes supermermercados Novo Mundo, Ganha Pouco e Mohmed e Companhia, proporcionam uma experiência de compras

tal que, que as tem como alternativa, independentemente das suas características, certamente tem maiores chances de sair com níveis de satisfação elevada – estudos posteriores podem vir a revelar a que se deve esta situação.

c) Para o Escalonamento Multidimensional (MDS), temos a concluir que, em termos de similaridade das variáveis não usadas na construção dos modelos acima descritos, temos a considerar:
   - os Supermercados que têm funcionários competentes tambem praticam preços acessíveis;
   - os supermercados que têm funcionários empenhados tambem têm atendimento rápido e são mais prestativos;
   - Se o supermercado é empenhado então, também possue atendimento personalizado;
   - Onde há empenho dos funcionários tambem há atenção dos funcionários e o comprometimento do supermercado;
   - A variedade dos serviços disponíveis está relacionado à qualidade de acesso ao supermercado;
   - E por fim , melhoria de qualidade está relacionado com melhoria de serviços.

Sendo assim, estes agrupamentos de variáveis indicam-nos as similaridades que as mesmas possuem e, a título exploratório, poderimos escolher e excluir variáveis redudantes de modo a facilitar as interpretação dos modelos que deles adviessem.

5.2 Recomendações

Tendo em conta as conclusões tiradas do estudo realizado sobre a base de dados *Satisfação com Supermercado Game ou Shoprite,* os investidores nas áreas de supermercados ou áreas similares, deviam replicar os supermercados com características de Shoprite em áreas que podessem captar a atenção de consumidores com características reveladas pelos modelos como, os que ganham até 500 Mtn, os solteiros e os que não têm nenhum nível de escolaridade. Segundo o relatório do Inquérito sobreo Orçamento Familiar (IOF) de 2022, publicado em Julho de 2023, na página 29, …, declara que *"Em geral, a maior parte de chefes de agregados familiares são camponeses (64,6%) seguidos de operários não agrícolas (9,2%). Por outro lado, nota-se que quase oito em cada dez (77,8%) mulheres chefes de agregados familiares são camponesas em comparação com a percentagem dos homens (59,6%)", "…maior parte das mulheres chefes de agregados familiares é divorciada/separada (36,4%)*

seguido de viuvez (34,5%) ," num outro desenvolvimento na página 63 declara que, *Em relação ao nível de escolaridade, observa-se que 90,7% da população sem nenhum nível de educação e 93,6% da que nunca frequentou escola, desenvolve as suas actividades económicas no ramo da agricultura, silvicultura e pesca."*

Sendo assim, concluimos que supermercados ou estabelecimentos que prestam serviços similares às de Shoprite, deviam ser implantados nas zonas rurais ou periurbanas.

Para os supermercados Game, dadas as características de potenciais clientes revelados pelos modelos, tendo em conta o relatório de IOF – 2022, os bens como Mobiliário, artigos de decoração, equipamento doméstico e manutenção corrente da habitação, que são bens mais fornecidos pelo supermercado Game,tem uma estrutura de consumo dos agregados familiares, com 14,5% na área rural e 19,7% na área urbana. Com isto entendemos que estes supermercados ou aos que tenham características similares, seria recomendável que os mesmos se implantassem no centro da cidade ou nas zonas periurbanas.

Relativamente ao nível de satisfação com os supermercados Game ou Shoprite, entendemos que, sendo a variável mais importante outro supermercado (*Outro_sup),* que em alguns casos chega a determinar o nível de satisfação sozinha, seria recomendável que os gestores destes supermercados fizessem um estudo periódico sobre o que os outros estão a oferecer a estes clientes (serviços ou productos), que de forma singular prevê a probabilidade de se alcançar o nível de satisfação mais alto.

6 Bibliografia

AGRESTI, A. *An Introduction to Categorical Data Analysis – 2ª Ed*, JohnWiley & Sons 2007

BOCHADO, A. O., CAETANO, J., COBRA, J., FONSECA, J. M., PORTUGAL, M., VARELA, M., BRANDÃO, N. G., FONTAN, O., CRUZ, R., MIRRANDA, S. *Desafios da Globalização, Marketing e Comunicação*, Escolar Editora 2013

CALSING, L. J., *Estudo sobre a satisfação do clientes da Fábrica de Móveis Klein Ltda*, UNIVATES, 2008

CASTRO, L. N., FERRRARI, D. G., *Introdução à Mineração de Dados, Conceitos Básicos, Algorítmos e Aplicações*, Saraiva 2016

FAUQUE, J. *Saber acolher os clientes*, Nathan París 1993

FÁVERO, L. P. & BELFIORE, E., *Manual de Análise de dados: Estatística e Modelagem Multivariada com EXCEL®, SPSS® e STATA®, Elsevier 2017*

GIL, A. C., *Métodos e Técnicas de Pesquisa social* – 6ª Ed, Atlas S. A. 2008

HAIR, J. F. ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. *Análise Multivariada de dados.* 5ª Ed., Porto Alegre: Bookman, 2005.

INE, *Inquérito Sobre o Orçamento Familiar 2022*, Relatório Final, Maputo 2023

JOÃO, ISABEL M. S *Aplicação da Regressão logística ordinal em estudos de lealdade de clientes. Evidência para a indústria hoteleira no Algarve. Revista Turismo & Turismo nº 17/18,* 2012

KOTLER, P. & KELLER, K, *Administração de Marketing – 14ª Ed*, Pearson Education 2012

KOTLER, P. & KELLER, K., *Administração de Marketing - 12ª Ed*, Pearson Education 2006

KRUSKAL, J. B. *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, 1964

MARCONI, M. A; LAKATOS, E. M. ***Técnicas de pesquisa****: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração e interpretação de dados* – 3 Ed. São Paulo: Atlas, 1996.

MATTAR, F. N. *Pesquisa de marketing* – 2 *Ed*, . São Paulo: Atlas, 1996

MORREIRA, I, *A excelência no Atendimento – 2ªEd*, Edições Técnicas 2010

MOURA, MARINA C. F. , *Diagnóstico No Modelo de Regressão ordinal, Dissertação de Mestrado, Instituto de Matemática e Estatística, da Universidade São Paulo, 2019*

REIS, E. *Estatística Multivariada Aplicada – 2ª Ed, Sílabo 2001*

SHAW, C. & IVENs, J. *Optimizar a Experiência do Cliente*, Europa-América de 2007

TRIVIÑOS A. *Introdução à Pesquisa em Ciências Sociais: A Pesquisa Qualitativa em Educação*. São Paulo: Ática 1987.

VILARES, M. J. & COELHO P. S. *Satisfação e Lealdade do Cliente – 2ª Ed*, Editora Escolar 2011

7 Anexos

7.1 Questionário

QUESTIONÁRIO DE PESQUISA

Caro Sr(a),este questionário é parte de um estudo que possui meramente um objectivo académico. A sua opinião é de extrema importância para a concretização desse trabalho. Todas as informações serão utilizadas de maneira agrupadas não sendo em nenhum momento identificado o respondeste.

Desde já, agradecemos pela colaboração na concretização deste trabalho, muito obrigada!

Nota: **Peço que responda a todas questões.**

DADOS PESSOAIS				
01. **Sexo:**	☐	Masculino	☐	Feminino
02. **Faixa Etária:**	☐	Menos de 15 anos	☐	Entre [31 à 40]anos
	☐	Entre [15 à 25]anos	☐	Entre [41 à 50]anos
	☐	Entre [26 à 30]anos	☐	Mais de 50 anos
03. **Estado Civil**	☐	Solteiro (a)	☐	Divorciado (a)
	☐	Casado (a)	☐	Viúvo (a)
04. **Nível de Escolaridade Concluído**				
	☐	Primário	☐	Superior
	☐	Básico	☐	Nenhum
	☐	Médio		
05. **Qual é o seu Rendimento Mensal?**				
	☐	Até 500 Mtn	☐	4500 Mtn á 10 000 Mtn
	☐	500 Mtn á 1500 Mtn	☐	10 000 Mtn á 15 000 Mtn
	☐	1500 á 3 000 Mtn	☐	Mais de 15 000 Mtn.
	☐	3000 á 45000 Mtn		
INFORMAÇÃO GERAL SOBRE O SERVIÇO DO SUPERMERCADO				
06.**Qual é o supermercado que frequenta?**				
	☐	Game	☐	Shoprite
07. **A quanto tempo é cliente do actual supermercado? (assinale com x o quadradinho)**				

☐ Menos de 1 ano

☐ Mais de 1 ano e menos de 2 anos

☐ Mais de 2 anos e menos de 3 anos

☐ Mais de 3 anos

08. **Com que frequência vem a este supermercado?**

☐ Diariamente

☐ Semanalmente

☐ Quinzenalmente

☐ Mensalmente

☐ Ocasionalmente

09. **Qual é o outro supermercado que costuma a frequentar? (cite pôr frequência)**

☐ Novo Mundo

☐ Ganha Pouco

☐ Mohmed e Companhia

☐ LM

☐ Luz

☐ Nosso Supermercado

Outro.....

10. **Como avalia a importância do Serviço de supermercado?**

☐ Não é importante ☐ Importante

☐ Pouco importante ☐ Muito importante

☐ Indiferente

Para as questões de 11à 27, Marcar com **X** a opção que melhor define o seu sentimento em relação as firmações, e o desempenho do seu supermercado.

TANGÍVEIS

11. **O supermercado possui vários tipos de serviço?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

12.**Como é que os produtos se encontram apresentados no supermercado?**

☐ Bem organizados ☐ Muito desorganizados

☐ Organizados ☐ Desorganizados

☐ Nem organizados e nem desorganizados

13. **A apresentação pessoal dos funcionários do supermercado é excelente?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

14. **O supermercado encontra-se em locais de fácil acesso?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

CONFIABILIDADE

15. **O supermercado cumpre sempre o que promete?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

16. **O que mais me faz comprar neste supermercado é:**

☐ Porque têm todos os produtos que necessito

☐ Porque fazem constantemente promoções

☐ Porque a qualidade dos produtos e boa

☐ Porque os preços baixos

☐ Outros. Especificar .....

17. **Na hora da escolha do Supermercado o que leva em conta? (marcar até 3 pôr ordem de importância).**

☐ Atendimento ☐ promoções/ofertas

☐ Ausência de filas no caixa ☐ qualidade

☐ Preço ☐ localização

☐ Serviços do caixa

☐ Variedade

18. **Sente-se satisfeito com os serviços prestados pôr este supermercado?**

☐ Pouco satisfeito ☐ Muito

☐ Pouco ☐ Muito satisfeito

☐ Mais ou Menos

19. **A qualidade dos produtos vendidos neste supermercado é melhor que nos outros supermercados?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

20. **Os preços praticados neste supermercado são mais baratos em relação a outros supermercados?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

21. **O supermercado mostra-se empenhado em facilitar mais a vida do cliente?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
PRESTEZA
22. **O atendimento do pessoal nos caixas do supermercado é sempre rápido?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
23. **Os funcionários do supermercado mostram-se sempre disponíveis para prestar o atendimento?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
24. **O número de funcionários do supermercado para o atendimento do cliente é adequado?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
25. **O supermercado têm-se esforçado bastante em melhorar a sua qualidade para fornecer melhores serviços?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
GARANTIA / SEGURANÇA
26. **Os funcionários do supermercado transmitem segurança, confiança aos seus clientes, pelo seu comportamento?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
27. **Os funcionários do supermercado têm competência e conhecimento suficiente para responder às necessidades do cliente?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
EMPATÍA
28. **Os funcionários do supermercado são atenciosos?** ☐ Discordo totalmente ☐ Concordo ☐ Discordo ☐ Concordo totalmente ☐ Mais ou Menos
29. **O meu supermercado têm funcionários que oferecem um atendimento personalizado (um bom atendimento) aos seus clientes?** ☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

30. **Os funcionários do supermercado resolvem as minhas preocupações e necessidades especificas sempre que entro em contacto com eles?**

☐ Discordo totalmente ☐ Concordo

☐ Discordo ☐ Concordo totalmente

☐ Mais ou Menos

Ordene de 1 à 5 os atributos a seguir conforme a importância que dás a cada item apresentado.

Nota: **1** é o que o (a) Sr (a) acha mais importante de todas e **5** é o que acha menos importante de todas

Não deve haver repetição de números e pôr favor não deixe sem preencher

Ordem	**Atributo**	
☐	**Tangíveis**	(Beleza das instalações e forma de apresentação dos funcionários)
☐	**Confiabilidade**	(Confiança no seu supermercado, fazer as coisas como prometem, satisfação dos serviços prestados, maior qualidade nos produtos e bons preços)
☐	**Presteza**	(Atendimento aos clientes com boa vontade)
☐	**Segurança**	(Conhecimento dos funcionários sobre os serviços)
☐	**Empatía**	(Consideração e atenção individualizada ao cliente)

Marque com X na coluna do supermercado que na sua opinião:

Atributos	**Game**	**Shoprite**	**Outros**	**São iguais**	**Não sabe**
	1	**2**	**3**	**4**	**5**
Possui os melhores serviços	☐	☐	☐	☐	☐
Possui o melhor atendimento	☐	☐	☐	☐	☐
Possui os melhores preços	☐	☐	☐	☐	☐
Possui melhor qualidade de produtos	☐	☐	☐	☐	☐
Preocupa-se com os clientes	☐	☐	☐	☐	☐

Se tiver alguma observação use o espaço a seguir

.....

.....

.....

8 Apêndice

8.1 Variáveis

Domínio (grupo de variáveis)	Variável - discrição
Dados ou características do indivíduo	Sexo – Sexo F_etaria – Faixa Etária E_civil – Estado Civil N_escolar – Nível de Escolaridade F_salarial – Rendimento Mensal
Informação geral sobre o serviço do supermercado	Sup_preferido – Nome do Supermercado de sua preferência Tempo_cliente – A quanto tempo e cliente Freq_sup – Com que frequência vem a este Supermercado Outro_sup – Qual é o outro Supermercado que costuma frequentar Importancia_sup – Importância do Supermercado
Tangíveis (Beleza das instalações e forma de apresentação dos funcionários)	Variedade_serviços– O Supermercado possui vários tipos de serviço Org_productos – Como encontram-se apresentados os produtos Aprumo_func – A apresentação dos funcionários e excelente Acesso_sup – Os locais dos Supermercados são de fácil acesso
Confiabilidade (Confiança no seu supermercado, fazer as coisas como prometem, Satisfação dos serviços prestados,	Compromentimento_sup – O Supermercado cumpre sempre o que promete Motivo_escolha – O que mais me faz comprar neste Supermercado MaisValiaNaEscolha – Na hora da escolha do Supermercado o que leva em conta Nível_satisfação – Sente-se satisfeito com este Supermercado

maior qualidade nos produtos e bons preços)	Melhoria_qualidade – A qualidade dos produtos é melhor que nos outros Supermercados Preço_acessivel – O preço praticado neste Supermercado é mais barato que nos outros Supermercados Empenho_sup – O Supermercado empenha-se em facilitar a vida das pessoas
Presteza (Atendimento aos clientes com boa vontade)	Ráp_atendimento – É rápido o atendimento dos caixas no Supermercado Empenho_func – Os funcionários mostram-se sempre disponíveis para o atendimento ao cliente NumStaff_adequado – É adequado o número de funcionários para o atendimento ao cliente Melhores_serv – O Supermercado tem-se esforçado em melhorar a sua qualidade para fornecer melhores serviços
Garantia / Segurança (Conhecimento dos funcionários sobre os serviços)	Func_compto – Os funcionários transmitem segurança, confiança aos seus clientes pelo seu comportamento Func_competentes – Os funcionários tem competência e conhecimento suficiente para responder às necessidades do cliente
Empatía (Consideração e atenção individualizada ao cliente)	Func_atenc – Os funcionários do supermercado são atenciosos At_perszado – Os funcionários oferecem um atendimento personalizado aos seus clientes Func_prestativos – Os funcionários resolvem as preocupações e necessidades especificam sempre que entro em contacto com eles
Ordem de valores (a que nível o indivíduo valoriza este atributo)	Tangíveis – Beleza das instalações e forma de apresentação dos funcionários Confiabilidade – Confiança, fazer coisas que prometem, satisfação nos serviços prestados, maior qualidade e bons preços Presteza – Atendimento ao cliente com boa vontade Segurança – Conhecimento dos funcionários sobre os serviços Empatia – Consideração e atenção individualizada ao cliente

1

8.2 Modelo de regressão logística binária

```
library(MASS)
library(stargazer)
library(car)
library(vcd)
#install.packages("readxl")
library(haven)
library(readxl)
library(dplyr)
library(caret)
library(pROC)
library(pacman)
#pacman::p_load(dplyr,psych,car,MASS, Destools,Quantpsic,ggplot2)

base_de_dados <- read_excel("base_de_dados.xlsx")
data<-base_de_dados

dados<-as.data.frame(data)
data$Sexo<-factor(data$Sexo)
data$F_et<-factor(data$F_et)
data$E_civ<-factor(data$E_civ)
data$N_esc<-factor(data$N_esc)
data$F_sal<-factor(data$F_sal)
data$Sup_preferido<-factor(data$Sup_preferido)

data$Sexo<-as.factor(data$Sexo)

data$Tangiveis<-factor(data$Tangiveis)
data$Confiabilidade<-factor(data$Confiabilidade)
data$Presteza<-factor(data$Presteza)
data$Seguranca<-factor(data$Seguranca)
data$Empatia<-factor(data$Empatia)

# Criar a nova base de dados usando a fun??o subset() para selecionar colunas
dados <- subset(data, select = c("Sexo","F_et","E_civ","N_esc","F_sal",
                                 "Sup_preferido","Tangiveis", "Presteza",
"Confiabilidade","Seguranca","Empatia"))
#categorias de referencia

dados$Sexo <- relevel(dados$Sexo, ref = "Feminino")
dados$Sup_preferido <- relevel(dados$Sup_preferido, ref = "0")
dados$E_civ <- relevel(dados$E_civ, ref = "solteiro")
```

```
dados$N_esc <- relevel(dados$N_esc, ref = "nenhum")
dados$F_et <- relevel(dados$F_et, ref = "menos de 15 anos")
dados$F_sal <- relevel(dados$F_sal, ref = "ate 500 Mtn")

levels(dados$Sup_preferido)
levels(dados$Sexo)
levels(dados$E_civ)
levels(dados$N_esc)
levels(dados$F_et)
levels(dados$F_sal)

set.seed(12345)
trainIndex <- createDataPartition(dados$Sup_preferido, p=0.8, list=FALSE)
data_train <- dados[trainIndex,]
data_test <- dados[-trainIndex,]
data_train<-as.data.frame(data_train)

ModeloInicial<-glm(Sup_preferido~., data=data_train, family = binomial (link =
"logit"))
car::vif(ModeloInicial)
ModeloInicial$rank
summary(ModeloInicial)
summary(stdres(ModeloInicial))


# Seleção de variáveis usando AIC
stepwise_model <- stepAIC(ModeloInicial, direction = "both")
summary(ModeloInicial)

# Seleção de variáveis usando BIC
stepwise_model_bic <- stepAIC(full_model, direction = "both", k = log(nrow(data)))
summary(stepwise_model_bic)


# Exibir valores dos outliers
print(outliers)


# Configuração da validação cruzada
control <- trainControl(method = "cv", number = 10)

# Treinamento do modelo com validação cruzada
```

```
cv_model <- train(Sup_preferido ~ ., data = dados, method = "glm", family =
binomial, trControl = control)
print(cv_model)


#verificacao de outlier
#plot(mod,wich=5)
#summary(stdres(ModeloInicial))

#plot(ModeloInicial,wich=5)

#Verificacao de multcolinearidade
#pairs.panels(dados)

summary(ModeloInicial)
Anova(ModeloInicial, type = "II")
Anova(ModeloInicial, type = "II",test="Wald")


Modelo2<-glm(Sup_preferido~Sexo+E_civ+N_esc+F_sal,data_train, family=binomial (link
= "logit"))
summary(Modelo2)

summary(stdres(Modelo2))

Anova(Modelo2, type = "II")
1-pchisq(Modelo2$null.deviance-Modelo2$deviance, Modelo2$df.null-
Modelo2$df.residual)

exp(coef(Modelo2))
exp(cbind(OR=coef(Modelo2),IC=confint.default(Modelo2)))

probs <-predict(Modelo2, data_train, type="response")
cm<-confusionMatrix(data=as.factor(ifelse(probs>=0.5,1,0)),
                    reference=data_train$Sup_preferido, positive="1")
print(cm)

curvaROC<-roc(data_train$Sup_preferido, probs)
curvaROC$auc
plot(curvaROC)

probs_test <-predict(Modelo2, data_test, type="response")
cm_test<-confusionMatrix(data=as.factor(ifelse(probs_test>=0.5,1,0)),
```

```
reference=data_test$Sup_preferido, positive="1")
cm_test$table

curvaROC_test<-roc(data_test$Sup_preferido,probs_test)
curvaROC_test$auc
plot(curvaROC)
plot(curvaROC_test, add=T, col=2)
```

8.3 Modelo de regressão logistica Ordinal

```
library(MASS)
library(stargazer)
library(caret)
library(car)
library(vcd)
library(readxl)

data <- read_excel("base_de_dadosOrd.xlsx")
data$Sexo<-factor(data$Sexo)
data$F_et<-factor(data$F_et)
data$E_civ<-factor(data$E_civ)
data$N_esc<-factor(data$N_esc)
data$F_sal<-factor(data$F_sal)
data$Sup_preferido<-factor(data$Sup_preferido)
data$Tempo_cliente<-factor(data$Tempo_cliente)
data$Freq_sup<-factor(data$Freq_sup)
data$Outro_sup<-factor(data$Outro_sup)
data$Importancia_sup<-factor(data$Importancia_sup)
data$Nivel_satisfacao<-factor(data$Nivel_satisfacao,
      levels = c("pouco satisfeito","pouco","mais ou menos","muito","muito
satisfeito"),ordered = TRUE)
data$Tangiveis<-factor(data$Tangiveis)
data$Confiabilidade<-factor(data$Confiabilidade)
data$Presteza<-factor(data$Presteza)
data$Seguranca<-factor(data$Seguranca)
data$Empatia<-factor(data$Empatia)

# Criar a nova base de dados usando a funcaoo subset() para selecionar colunas
dados <- as.data.frame(subset(data, select =
c("Sexo","F_et","E_civ","N_esc","F_sal","Sup_preferido",
```

```
                                "Nivel_satisfacao"
,"Tempo_cliente","Freq_sup","Outro_sup","Importancia_sup",
                                 "Tangiveis", "Presteza",
"Confiabilidade","Seguranca","Empatia")))

View(dados)
table(dados$Nivel_satisfacao)/nrow(dados)

str(dados)
#categorias de referencia
dados$Sexo <- relevel(dados$Sexo, ref = "Feminino")
dados$E_civ <- relevel(dados$E_civ, ref = "solteiro")
dados$N_esc <- relevel(dados$N_esc, ref = "nenhum")
dados$F_sal <- relevel(dados$F_sal, ref = "ate 500 Mtn")
#dados$Nivel_satisfacao <- relevel(dados$Nivel_satisfacao, ref = "pouco
satisfeito")

levels(dados$Sexo)
levels(dados$E_civ)
levels(dados$N_esc)
levels(dados$F_sal)
levels(dados$Outro_sup)
levels(dados$Confiabilidade)
levels(dados$Nivel_satisfacao)

data2 <- read_excel("base dedados Cod.xlsx")
data2<- data2[, -c(11:17,19:30)]
View(data2)
 dados2<-as.data.frame(data2)

 modelo<-lm(Nivel_satisfacao~.,data = dados2)
 library(car)
 vif(modelo)

library(DescTools)

matrizCor<- cor(dados2, method = "kendall") #Kendal tau-beta
correlacao_spearman<-cor(x=as.matrix(dados2),method="spearman")


# Ajustar um modelo auxiliar com glm para obter os resíduos
modelo_glm <- glm(Nivel_satisfacao ~ ., data = dados, family = binomial(link =
"logit"))
```

```
residuos <- residuals(modelo_glm, type = "pearson")

# Testar a autocorrelação nos resíduos
#resultado_teste <- dwtest(residuos ~ 1)
#print(resultado_teste)
library(lmtest)
resultado_DW<-dwtest(modelo_glm, alternative = "less")
print(resultado_DW)


set.seed(12345)
trainIndex <- createDataPartition(dados$Nivel_satisfacao, p=0.8, list=FALSE)
data_train <- dados[trainIndex,]
data_test <- dados[-trainIndex,]

modeloAux<-polr(Nivel_satisfacao ~ 1, data=data_train)
aux<-c(rep(0,sum(sapply(data_train[-7], nlevels))-
ncol(data_train)+1),modeloAux$zeta)
modeloInicial<-polr(Nivel_satisfacao ~ ., data=data_train, start=aux)

summary(modeloInicial)
car::poTest(modeloInicial)

# Teste de Wald para obter valores-p
wald_results <- summary(modeloInicial)$coefficients[, c("Value", "Std. Error")]
wald_results <- cbind(wald_results, p_value = (1 - pnorm(abs(wald_results[,
"Value"]) / wald_results[, "Std. Error"])) * 2)

# Visualizar os resultados
print(wald_results)


# Resumo do modelo com valores-p
summary(modeloInicial, test = "Chisq")

summary(modeloInicial)

typeII<-data.frame()
for (i in c(1:6,8:ncol(data_train))){
  auxSin<-c(rep(0,modeloInicial$edf-length(modeloInicial$zeta)-
nlevels(data_train[,i])+1),modeloInicial$zeta)
  modeloSin<-polr(Nivel_satisfacao ~ ., data=data_train[,-i], start=auxSin)
```

```
  LRChisq<-modeloSin$deviance-modeloInicial$deviance
  Df<-modeloInicial$edf-modeloSin$edf
  pVal<-1-pchisq(LRChisq, Df)
  typeII<-rbind(typeII,data.frame(var=names(data_train)[i],LRChisq,Df,pVal))
}
typeII

dados2 <- as.data.frame(subset(data, select = c("Nivel_satisfacao"
,"Sexo","E_civ","F_sal","Outro_sup")))
View(dados2)

set.seed(12345)
trainIndex <- createDataPartition(dados2$Nivel_satisfacao, p=0.8, list=FALSE)
data_train <- dados2[trainIndex,]
data_test <- dados2[-trainIndex,]

modeloAux2<-polr(Nivel_satisfacao ~ 1, data=data_train)
aux<-c(rep(0,sum(sapply(data_train[-1], nlevels))-
ncol(data_train)+1),modeloAux$zeta)
modelo2<-polr(Nivel_satisfacao ~Sexo+E_civ+F_sal+Outro_sup, data=data_train,
start=aux)

# Teste de Wald para obter valores-p
wald_results <- summary(modelo2)$coefficients[, c("Value", "Std. Error")]
wald_results <- cbind(wald_results, p_value = (1 - pnorm(abs(wald_results[,
"Value"]) / wald_results[, "Std. Error"])) * 2)
# Visualizar os resultados
print(wald_results)
```

8.4 Árvore de classificação binária

```
library(MASS)
library(stargazer)
library(caret)
library(car)
library(vcd)
library(readxl)
library(rpart)
library(rpart.plot)
```

```
library(ggplot2)

setwd("C:/Users/Salomao/Desktop/Projecto de Dissertassao/Regressao Ordinal")
data <- read_excel("base_de_dados.xlsx")

str(data)
summary(data)

# Converte as variáveis para fatores

data$Sexo <- factor(data$Sexo)
data$F_et <-factor(data$F_et, levels = c("menos de 15 anos","entre 15 a 25
anos","entre 26 a 30 anos",
                                         "entre 31 a 40 anos","entre 41 a 50 anos","mais
de 50 anos"),ordered = TRUE)
data$E_civ <- factor(data$E_civ)
data$N_esc <- factor(data$N_esc,levels =
c("nenhum","primario","basico","medio","superior"),ordered = TRUE)
data$F_sal <- factor(data$F_sal,levels = c("ate 500 Mtn","500 a 1500 Mtn","1500 a
3000 Mtn","3000 a 4500 Mtn",
                                       "4500 a 10 000 Mtn","10 000 a 15 000 Mtn","mais
de 15 000 Mtn"),ordered = TRUE)
data$Sup_preferido <- factor(data$Sup_preferido)

data$Tangiveis <- factor(data$Tangiveis,levels = c("nao importante","menos
importante","mais ou menos","importante",
                                         "mais importante"),ordered = TRUE)
data$Confiabilidade <- factor(data$Confiabilidade,levels = c("nao
importante","menos importante","mais ou menos","importante",
                                         "mais importante"),ordered = TRUE)
data$Presteza <- factor(data$Presteza,levels = c("nao importante","menos
importante","mais ou menos","importante",
                                         "mais importante"),ordered = TRUE)
data$Seguranca <- factor(data$Seguranca,levels = c("nao e importante","menos
importante","mais ou menos","importante",
                                         "mais importante"),ordered = TRUE)
data$Empatia <- factor(data$Empatia,levels = c("nao e importante","menos
importante","mais ou menos","importante",
                                         "mais importante"),ordered = TRUE)

# Criar a nova base de dados usando a fun??o subset() para selecionar colunas
```

```
dados <- subset(data, select =
c("Sexo","F_et","F_sal","E_civ","N_esc","Sup_preferido","Empatia",

"Tangiveis","Seguranca","Presteza","Confiabilidade"))

View(dados)
#categorias de referencia
levels(dados$Sup_preferido)
levels(dados$Sexo)
levels(dados$E_civ)
levels(dados$N_esc)
levels(dados$F_et)
levels(dados$F_sal)


set.seed(12345)
trainIndex <- createDataPartition(dados$Sup_preferido, p=0.8, list=FALSE)
data_clasif_train <- dados[trainIndex,]
data_clasif_test <- dados[-trainIndex,]

#usando o criterio de Entropia
set.seed(12345)
modeloEntropia<-rpart(Sup_preferido~., data=data_clasif_train, method = "class",
                  minbucket=ceiling(0.02*nrow(data_clasif_train)),cp=0,
                  parms=list(split="information"),maxsurrogate = 0)
sum(modeloEntropia$frame$var == "<leaf>")

rpart.plot(modeloEntropia,type=5,extra = 105,nn=TRUE,tweak=1.0,
           main="Árvore de classificação - Binária")


asRules(modeloEntropia)
summary(modeloEntropia)

probs <-predict(modeloEntropia,data_clasif_test,type="prob")
cm<-confusionMatrix(data=as.factor(ifelse(probs[,2]>=0.5,1,0)),
                    reference=data_clasif_test$Sup_preferido,positive="1")
cm$table
cm$overall[1:2]
cm$byClass[1:2]


#importancia das variaveis
```

```
modeloEntropia$variable.importance

barplot(modeloEntropia$variable.importance.importance/sum(modeloEntropia$variable.i
mportance))

dado <- modeloEntropia$variable.importance

# Calcular porcentagens
percentagens <- dado / sum(dado) * 100
print(percentagens)

# Criar o gráfico de barras com barplot
grafico <- barplot(percentagens, col = "skyblue",ylim = c(0, 100), main = "Gráfico
de Importancia/rank de Variaveis")

# Adicionar as porcentagens nas barras
text(grafico, dado, labels = sprintf("%.1f%%", percentagens), pos = 3, col =
"black")



modeloEntropia$cptable
plotcp(modeloEntropia)

modeloEntropiaPodado <- prune(modeloEntropia, cp =  0.0226)

rpart.plot(modeloEntropiaPodado,type=2, extra=105,nn=TRUE,tweak=1.0,main="Árvore de
classificação - Binária (Podada)")

dataVCR<-data_clasif_train
dataVCR$Sup_preferido<-factor(dataVCR$Sup_preferido,
levels=rev(levels(dataVCR$Sup_preferido)),
                           labels=c("Game","Shoprite"))
table(dataVCR$Sup_preferido)/nrow(dataVCR)


set.seed(1712)
# Certifique-se de que 'Sup_preferido' é a variável resposta na 6ª coluna
if (colnames(dataVCR)[6] != "Sup_preferido") {
  stop("Certifique-se de que 'Sup_preferido' é a variável na 6ª coluna de
'dataVCR'.")
}
```

```
# Certifique-se de que 'modeloEntropia' está definido corretamente
if (!exists("modeloEntropia") || !("cptable" %in% names(modeloEntropia))) {
  stop("Certifique-se de que 'modeloEntropia' está definido corretamente e contém
'cptable'.")
}

# Definir os parâmetros de treino
vcr <- train(
  y = dataVCR$Sup_preferido,
  x = dataVCR[,-6],  # Exclui a 6ª coluna (variável resposta)
  method = "rpart",
  control = rpart.control(minbucket = ceiling(0.01 * nrow(dataVCR))),
  metric = "AUC",
  parms = list(split = 'information'),
  tuneGrid = expand.grid(cp = modeloEntropia$cptable[,1]),
  trControl = trainControl(
    method = "repeatedcv",
    number = 5,
    repeats = 20,
    summaryFunction = multiClassSummary,
    classProbs = TRUE,
    returnResamp = "all",
    savePredictions = TRUE
  )
)

# Ver os resultados do modelo
print(vcr)

# Extrair os resultados da validação cruzada
results <- vcr$results

# Encontrar o índice da linha com o melhor desempenho (maior AUC)
best_row <- which.max(results$AUC)

# Obter o valor de cp correspondente ao melhor desempenho
best_cp <- results$cp[best_row+1]
# tomamos a terceira linha dado o
#principio de parsimonia, uma vez que as arvores sao quase iguais

modeloEntropiaFinal <- prune(modeloEntropia, cp = best_cp)
rpart.plot(modeloEntropiaFinal,type=5,extra = ,nn=TRUE,tweak=1.0,
           main='Árvore de Classificação Final')
```

8.5 Árvore de classificação Ordinal

```
# Load the MLmetrics package
library(MLmetrics)

# Carregue as bibliotecas necessárias
library(rpart)
library(rpart.plot)
library(arules)
library(arulesViz)
library(caTools)
library(caret)
library(readxl)

# Leitura dos dados

setwd("C:/Users/Salomao/Desktop/Projecto de Dissertassao/Regressao Ordinal")
data <- as.data.frame(read_excel("base_de_dadosOrd.xlsx"))

# Converte as variáveis para fatores
data$Sexo <- factor(data$Sexo)
data$F_et <- factor(data$F_et)
data$E_civ <- factor(data$E_civ)
data$N_esc <- factor(data$N_esc)
data$F_sal <- factor(data$F_sal)
data$Sup_preferido <- factor(data$Sup_preferido)
data$Tempo_cliente <- factor(data$Tempo_cliente)
data$Freq_sup <- factor(data$Freq_sup)
data$Outro_sup <- factor(data$Outro_sup)
data$Importancia_sup <- factor(data$Importancia_sup)
data$Nivel_satisfacao <- factor(data$Nivel_satisfacao,
                                levels = c("pouco satisfeito", "pouco", "mais ou
menos", "muito", "muito satisfeito"),
                                ordered = TRUE)
data$Tangiveis <- factor(data$Tangiveis)
data$Confiabilidade <- factor(data$Confiabilidade)
data$Presteza <- factor(data$Presteza)
data$Seguranca <- factor(data$Seguranca)
data$Empatia <- factor(data$Empatia)
```

```
dados <- data[,-c(11:17, 19:30)]

# Visualize os dados
View(dados)

levels(dados$Nivel_satisfacao)
# Defina a semente para reprodutibilidade
set.seed(12345)

# Divida os dados em conjunto de treinamento e teste
trainIndex <- createDataPartition(dados$Nivel_satisfacao, p = 0.8, list = FALSE)
data_train <- dados[trainIndex, ]
data_test <- dados[-trainIndex, ]

# Ajuste o modelo de árvore de decisão usando rpart
modeloEntropiaOrd <- rpart(Nivel_satisfacao ~ ., data = data_train, method =
"class",
                           minbucket = ceiling(0.02 * nrow(data_train)), cp = 0,
                           parms = list(split = "information"), maxsurrogate = 0)

# Resumo do modelo
summary(modeloEntropiaOrd)


# Visualização da árvore de decisão
rpart.plot(modeloEntropiaOrd,type=5, extra=100, main = "Árvore de classificação -
Ordinal ", nn = TRUE, tweak = 1.2)
sum(modeloEntropiaOrd$frame$var == "<leaf>")   # conta o numero de folhas da arvore
max(rpart:::tree.depth(as.numeric(rownames(modeloEntropiaOrd$frame)))) #conta a
profundidade

# Realize previsões no conjunto de teste
probs <- predict(modeloEntropiaOrd, data_test, type = "class")

# Crie a matriz de confusão
cm <- confusionMatrix(data = as.factor(probs), reference =
data_test$Nivel_satisfacao)
print(cm)

cm$table
cm$overall[1:2]
cm$byClass[1:2]
```

```
#importancia das variaveis
modeloEntropiaOrd$variable.importance

barplot(modeloEntropiaOrd$variable.importance/sum(modeloEntropiaOrd$variable.import
ance))

# Exemplo de dados
dado <- modeloEntropiaOrd$variable.importance

# Calcular porcentagens
percentagens <- dado / sum(dado) * 100
print(percentagens)

# Criar o gráfico de barras com barplot
grafico <- barplot(percentagens, col = "skyblue",ylim = c(0, 100), main = "Gráfico
de Importancia/rank de Variaveis")

# Adicionar as porcentagens nas barras
text(grafico, dado, labels = sprintf("%.1f%%", percentagens), pos = 3, col =
"black")



#poda da arvore
modeloEntropiaOrd$cptable

plotcp(modeloEntropiaOrd)

#arvore podada
modeloEntropiaOrdPodado <- prune(modeloEntropiaOrd, cp =
modeloEntropiaOrd$cptable[4,1])

#Indica-se a linha 5 porque é onde está localizada a subárvore desejada
rpart.plot(modeloEntropiaOrdPodado,main = "Árvore de Classificação - Ordinal
(Podada)",
           type=5,extra = 100,nn=TRUE,tweak=1.0)

sum(modeloEntropiaOrdPodado$frame$var == "<leaf>")   # conta o numero de folhas da
arvore podada
max(rpart:::tree.depth(as.numeric(rownames(modeloEntropiaOrdPodado$frame)))) #conta
a profundidade

#Validacao Cruzada
```

```
dataVCR<-data_train
dataVCR$Nivel_satisfacao<-factor(dataVCR$Nivel_satisfacao,
levels=rev(levels(dataVCR$Nivel_satisfacao)),
                     labels=c("pouco satisfeito", "pouco", "mais ou menos", "muito",
"muito satisfeito"))
table(dataVCR$Nivel_satisfacao)/nrow(dataVCR)
levels(dataVCR$Nivel_satisfacao)
levels(dataVCR$Nivel_satisfacao) <- make.names(levels(dataVCR$Nivel_satisfacao),
unique = TRUE)
levels(dataVCR$Nivel_satisfacao)

set.seed(1712)
vcr<-train(y=dataVCR$Nivel_satisfacao, x=dataVCR[,-11],
           method = "rpart",
           control = rpart.control(minbucket=ceiling(0.01*nrow(dataVCR))),
           metric="AUC", parms=list(split='information'),
           tuneGrid = expand.grid(cp=modeloEntropiaOrd$cptable[,1]),
           trControl = trainControl(method="repeatedcv", number=5, repeats=20,
                                    summaryFunction=multiClassSummary,
                                    classProbs=TRUE, returnResamp = "all",
                                    savePredictions = TRUE))

modeloEntropiaOrd$cptable #lo imprimimos para ver la equivalencia entre alpha y n.
de hojas


# Extrair os resultados da validação cruzada
results <- vcr$results

# Encontrar o índice da linha com o melhor desempenho (maior AUC)
best_row <- which.max(results$AUC)

# Obter o valor de cp correspondente ao melhor desempenho
best_cp <- results$cp[best_row]

modeloEntropiaOrdFinal <- prune(modeloEntropiaOrd, cp = best_cp)
rpart.plot(modeloEntropiaOrdFinal,type=2, extra=105,nn=TRUE,tweak=1.0,
           main='Árvore de Classificação - Ordinal (Final)')
# Exemplo de dados
dados <- modeloEntropiaOrdFinal$variable.importance

# Calcular percentagens
percentagens <- dados / sum(dados) * 100
```

```
print(percentagens)

# Criar o gráfico de barras com barplot
graficoF <- barplot(percentagens, col = "skyblue",ylim = c(0, 100), main = "Gráfico
de Importancia/rank de Variaveis")

# Adicionar as porcentagens nas barras
text(graficoF, dados, labels = sprintf("%.1f%%", percentagens), pos = 3, col =
"black")



# Realize previsões no conjunto de teste para arvore final
probsf <- predict(modeloEntropiaOrdFinal, data_test, type = "class")

# Crie a matriz de confusão para arvore final
cmf <- confusionMatrix(data = as.factor(probs), reference =
data_test$Nivel_satisfacao)
print(cmf)

cmf$table
cmf$overall[1:2]
cmf$byClass[1:2]

boxplot(AUC~cp,data=vcr$resample,xlab="")
boxplot(Kappa~cp,data=vcr$resample,xlab="")

library(rpart)
library(rpart.plot)
#indico a fila 4 pois e donde se poda a subárvore desejada
modeloEntropiaOrdPodado <- prune(modeloEntropiaOrd, cp =
modeloEntropiaOrd$cptable[5,1])
rpart.plot(modeloEntropiaOrdPodado,main = "Árvore de Classificação Podada (final)",
extra = 100,nn=TRUE,tweak=1.2)

modeloEntropiaOrdPodado$variable.importance

library(pROC)
roc(data_test$Nivel_satisfacao,
    predict(modeloEntropiaOrdPodado,data_test,type="prob")[,2], direction="<")$auc

freq_rel<-prop.table(table(data_test$Nivel_satisfacao))
```

```
probs <-predict(modeloEntropiaOrdPodado, data_test, type="prob")
cm<-confusionMatrix(data=as.factor(ifelse(probs[,2]>=0.5,1,0)),
                    reference=data_test$Nivel_satisfacao, positive="1")
cm$table
cm$overall[1:2]
cm$byClass[1:2]

cm2<-confusionMatrix(data=as.factor(ifelse(probs[,2]>=0.73,1,0)),
                     reference=data_test$Sobrepeso, positive="1")
cm2$table
cm2$overall[1:2]
cm2$byClass[1:2]
```

8.6 MDS

```
library(smacof)
library(readxl)
library(data.table)
library(DescTools)
library(questionr)
library(OneR)

setwd("C:/Users/Salomao/Desktop/Projecto de Dissertassao/Regressao Ordinal")

data <- read_excel("base dedados Cod.xlsx")
data<- data[, -c(1:10,16:18,31:35)]
str(data)
View(data)


dados<- as.data.frame(data)
#head(dados)

matrizCor<- cor(dados, method = "kendall") #Kendal tau-beta
matrizDisim <- sim2diss(matrizCor, method = 1)
MDSvariables_V<- mds(matrizDisim, type = "ordinal")
MDSvariables_V

set.seed(123)
permCol <- permtest(MDSvariables_V, nrep = 500, verbose = FALSE)
permCol
```

```
plot(permCol)

#fitbestdados <- MDSvariables_V
#set.seed(12345)
#for(i in 1:100) {
#  fitran <- mds(matrizDisim, type = "ordinal", init = "random")
#  if ((fitran$stress < fitbestdados$stress) & (cor(fitran$delta,fitran$dhat)>0)) {
#    fitbestdados <- fitran
#  }
#}
#fitbestdados

#par(cex = 1.3)
#plot(fitbestdados, plot.type = "stressplot",main='Decomposição de Stress')
#abline(h=100/fitbestdados$nobj, lty=2)


#plot(fitbestdados,main=' Gráfico de configuração')


library(ClustOfVar)
tree <- hclustvar(X.quanti = dados, X.quali =NULL)
plot(tree)

particion<-cutreevar(tree, 3)
df_aux<-data.frame(x=(fitbestdados$conf)[,1],
                   y=(fitbestdados$conf)[,2],
                   cluster=particion$cluster)
rownames(df_aux)<-rownames(fitbestdados$conf)
ggplot(df_aux, aes(x=x, y=y, col=as.factor(cluster))) + xlim (c(-1,1)) +
  geom_text(label=rownames(df_aux)) + theme_bw() + coord_fixed() +
  theme(legend.position = "none")

set.seed(12345)
stab <- stability(tree, B = 50, graph = F)
boxplot(stab$matCR, main = "Estabilidade dos diferentes agrupamentos")

particion<-cutreevar(tree, 7)
df_aux<-data.frame(x=(fitbestdados$conf)[,1],
                   y=(fitbestdados$conf)[,2],
                   cluster=particion$cluster)
rownames(df_aux)<-rownames(fitbestdados$conf)
```

```
ggplot(df_aux, aes(x=x, y=y, col=as.factor(cluster))) + xlim (c(-1,1)) +
  geom_text(label=rownames(df_aux)) + theme_bw() + coord_fixed() +
  theme(legend.position = "none")

par(mfrow = c(3, 4))
boxplot(dados$Func_competentes~dados$Preco_acessivel)
boxplot(dados$Func_compto~dados$Compromentimento_sup )
boxplot(dados$Func_compto  ~dados$Func_atenc)
boxplot(dados$Compromentimento_sup  ~dados$Func_atenc )
boxplot(dados$Melhoria_qualidade  ~dados$Melhores_serv )
boxplot(dados$Rap_atendimento  ~dados$Func_prestativos )
boxplot(dados$Rap_atendimento  ~dados$Empenho_func )
boxplot(dados$Func_prestativos  ~dados$Empenho_func )
boxplot(dados$Variedade_serviços  ~dados$Acesso_sup )
boxplot(dados$Aprumo_func  ~dados$NumStaff_adequado )
boxplot(dados$At_perszado  ~dados$Empenho_sup )
```