MARCELO CARLOS RIBEIRO

UMA ABORDAGEM PARA CLASSIFICAÇÃO MONOTÔNICA DE DADOS CORRELACIONADOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Fernando Luiz Pereira de Oliveira

Coorientador: Tiago Martins Pereira

VIÇOSA - MINAS GERAIS

2019

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

R484a
2019

Ribeiro, Marcelo Carlos, 1987-
Uma abordagem para classificação monotônica de dados correlacionados / Marcelo Carlos Ribeiro. - Viçosa, MG, 2019.
72 f. : il. ; 29 cm.

Orientador: Fernando Luiz Pereira de Oliveira.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 67-72.

1. Classificação e seleção (Estatística). 2. Variáveis (Matemática). 3. Correlação (Estatística) . I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.54

MARCELO CARLOS RIBEIRO

UMA ABORDAGEM PARA CLASSIFICAÇÃO MONOTÔNICA DE DADOS CORRELACIONADOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 18 de dezembro de 2019.

Assentimento:

Marcelo Carlos Ribeiro
Autor

Fernando Luiz Pereira de Oliveira
Orientador

A minha filha,

Sophia.
DEDICO

AGRADECIMENTOS

A Deus por me conceder a esperança, força e resiliência necessária para que eu pudesse chegar ao final dessa etapa, que sem sombra de dúvidas, considero como a realização de um sonho.

Aos meus pais, por toda dedicação, esforço e apoio demonstrados em todas as etapas da minha vida. Certamente, eles são os responsáveis todas as minhas conquistas.

À minha esposa e grande amiga Giselle, por todo o seu apoio, companheirismo e paciência. Muito obrigado por me ajudar a realizar esse sonho.

À Universidade Federal de Viçosa, especialmente ao programa de pós-graduação em Estatística Aplicada e Biometria da UFV, por todo aprendizado proporcionado.

Ao meu orientador, professor Dr. Fernando Luiz Pereira de Oliveira e coorientador, professor Dr. Tiago Martins Pereira que tiveram por mim, respeito confiança e compreensão.

Aos professores, Dr. Anderson Castro Soares de Oliveira, Dr. Antônio Policarpo Souza e Dra. Graziela Dutra Rocha Gouvea, por todas as contribuições fornecidas para a melhoria do trabalho.

Aos demais professores que compõem o departamento de estatística (DET) da Universidade Federal de Viçosa, por toda atenção e aprendizado transmitido.

Aos secretários Anita e Júnior, por toda atenção, profissionalismo e eficiência.

Ao departamento de estatística da Universidade de Ouro Preto (UFOP), por apoiar e incentivar de todas as maneiras possíveis o meu afastamento para o término desse doutorado.

Aos meus amigos, por me motivarem em diversas etapas dessa caminhada.Arthurzão, Marianam, André, Álisson, Alex, Ana Carolina, Ananda, Anderson Teodoro, Ben Deivide, Brunão, Clebis, Diegão (The legend), Matheus, Micherlânia, Marquinhos, Nilze, Rafael, Josino (Josinight), Gabriel, Gabriele, Ramon, Rosita, Helgem,Tiago, Didi, Spencer, cosminha, Juliano, Rivert, Maurício, Alex, Rosane, Lidiane, e tantos outros que por hora não lembro o nome, mas sempre serei grato.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

RIBEIRO, Marcelo Carlos, D.Sc., Universidade Federal de Viçosa, dezembro de 2019. **Uma abordagem para a classificação monotônica de dados correlacionados**. Orientador: Fernando Luiz Pereira de Oliveira. Coorientador: Tiago Martins Pereira.

A classificação ordenada está cada vez mais atraindo o interesse de áreas como estatística, ciências da computação e pesquisa operacional. A restrição de monotonicidade indica uma relação entre o rótulo da classe com uma ou mais variáveis (atributos). Nesta tese, apresentam-se duas contribuições resultantes de um trabalho de investigação sobre a classificação monotônica de dados correlacionados. Uma consiste em propor uma metodologia que se baseia no método CPP-tri proposto por Sant'Anna, Costa e Pereira (2015), que considere a correlação entre os atributos no cálculo da probabilidade do indivíduo pertencer a classe. A outra, consiste em fornecer um pacote R para o método proposto, denominado como CPP-cor Ribeiro et al. (2020). Os algoritmos desenvolvidos basearam-se no código em R disponível em Silva (2016). A metodologia proposta não só agrega a informação relacionada à correlação das variáveis ao método, como apresenta resultados significativamente superiores quando comparados aos resultados obtidos pela metodologia tradicional, o método CPP-tri.

Palavras-chave: Classificação monotônica. Múltiplas variáveis correlacionadas.

ABSTRACT

RIBEIRO, Marcelo Carlos, D.Sc., Universidade Federal de Viçosa, December, 2019. **An approach to the monotonic classification of correlated data**. Adviser: Fernando Luiz Pereira de Oliveira. Co-adviser: Tiago Martins Pereira.

The classification ordered is increasingly attracting interest from areas such as statistics, computer science and operational research. The monotonicity constraint indicates a relationship between the class label with one or more variables (attributes). In this thesis, two contributions resulting from research work on the monotonic classification of correlated data are presented. One is to propose a methodology based on the CPP-tri method proposed by Sant'Anna, Costa e Pereira (2015), which considers the correlation between attributes when calculating the probability of the individual belonging to the class. The other is to provide an R package for the proposed method, called CPP-cor Ribeiro et al. (2020). The developed algorithms were based on the R code available at Silva (2016). The proposed methodology not only aggregates information related to the correlation of variables to the method, but also presents significantly superior results when compared to the results obtained by the traditional methodology, the CPP-tri method.

Keywords: Monotonic classification. Multiple correlated variables.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Página

1 INTRODUÇÃO GERAL

1.1 Motivação

A classificação é uma tarefa inata ao ser humano e indispensável na maioria das suas atividades. O seu uso está associado a necessidade em dispor elementos em grupos ou classes de acordo com o seu grau de semelhança, de forma que, elementos pertencentes a mesma classe apresentam semelhança no que diz respeito as suas características observadas; enquanto que, elementos de classes distintas dispõem de características diferentes. Além disso, a classificação possibilita estabelecer ordem e controle em problemas complexos e muitas vezes desorganizados, identificando elementos próximos e reduzindo um número aparentemente imensurável de elementos individuais a um número mais gerenciável de classes logicamente relacionadas.

Se diz que uma característica apresenta natureza ordinal quando existe uma ordenação natural de seus elementos, evidenciando intensidades crescentes ou decrescentes. Este tipo de classificação apresenta por si só um importante papel em diversos campos da ciência e da sociedade. Desde tempos imemoriais a humanidade busca pela hierarquia e pela ordenação, seja em aplicações rudimentares como a separação de animais mais e menos produtivos nas primeiras criações, na escolha de lideranças locais por meio de disputas de força, destreza e inteligência, dentre outras abordagens.

Mesmo no cotidiano existem momentos nos quais a classificação ordenada ocorre de forma natural, como por exemplo, na seleção de restaurantes favoritos, criação de *playlists* de músicas e vídeos, compra de bens de consumo, etc. Muitas destas aplicações cotidianas hoje em dia representam desafios para diversas instituições, que tem como objetivo prever, por exemplo, o comportamento do consumidor e oferecer a ele o produto que está melhor colocado em sua lista de desejos. Aplicações desta natureza utilizando inteligência artificial e aprendizado de máquina, bem como, outras modelagens como *credit scoring* e análise de risco, são aplicações sofisticadas de metodologias de classificação ordenada.

Em um estudo de classificação, procura-se desenvolver um classificador preciso ou descobrir a estrutura preditiva do problema em questão. Esse último em es-

pecífico, consiste em fornecer caracterizações simples do relacionamento entre as variáveis que determinam quando um indivíduo está em uma classe e não em outra.

Os avanços computacionais vivenciados desde as últimas décadas permitem que diversos campos da ciência trabalhem de forma transversal na exploração e desenvolvimento de métodos. Entre as tentativas realizadas, pode-se mencionar estatística, inteligência artificial e pesquisa operacional. Os métodos propostos nessas áreas, exigem para sua construção uma estrutura realista que seja capaz de acomodar a natureza multivariada de problemas frequentes no mundo real.

Apesar dos avanços da ciência no que diz respeito a construção de métodos de classificação de dados ordenados, as metodologias clássicas desconsideram a presença de correlação entre as variáveis que descrevem cada indivíduo, e justamente deste aspecto emerge a contribuição dessa tese, pois, a metodologia aqui proposta considera não só os vetores de médias para a identificação das classes, como também, a matriz de correlação entre as variáveis que descrevem cada elemento. Desse modo, o intuito é aprimorar a acurácia de uma das metodologias tradicionais utilizadas na solução do problema de classificação ordenada monotônica, o método da composição probabilística de preferências em modo tricotômico (CPP-tri).

De acordo com o estado da arte, o método de classificação CPP-tri e suas variações, não tem considerado a correlação entre as variáveis no cálculo da probabilidade de cada indivíduo estar acima ou abaixo de cada classe. Sendo assim, essa tese detém duas perspectivas quanto a investigação de uma abordagem que atue examinando a correlação entre as variáveis; a primeira consiste em melhorar a descrição e identificação de cada classe presente no problema, considerando a matriz de correlação como um perfil de referência das classes. Por consequência, a segunda resulta em melhorar a conformidade na classificação dos indivíduos às suas classes de origem, com base apenas nas características observadas e apresentadas por eles.

1.2 Objetivos

Resumimos abaixo as contribuições desta tese para o avanço da metodologia de classificação de dados ordinais. Nesta tese, tivemos as seguintes contribuições:

1. apresentar uma revisão sistemática sobre métodos de classificação e metodologias de classificação ordinal;

2. criar, com base na metodologia CPP-tri, um procedimento para automatizar a criação dos perfis de referência e adequação do banco de dados para a classificação ordinal;
3. introduzir a correlação como parâmetro na classificação ordinal de dados por meio da distribuição conjunta de probabilidades;
4. apresentar uma biblioteca em linguagem R, de modo a facilitar o uso por pesquisadores interessados em solucionar problemas de classificação ordenada considerando ou desprezando o relacionamento entre as variáveis;
5. facilitar a utilização da metodologia por usuários leigos na linguagem R, por meio de uma interface gráfica amigável.
6. criação de um método que permite levar em conta as correlações entre as variáveis, tal fato abre frente para a utilização da metodologia de classificação em conjunto de dados que não supõe independência, o que permite a aplicação da mesma em outros cenários, concedendo maior abrangência para utilização da metodologia em outras áreas. Do ponto de vista da aplicação, isto faz-se interessante, pois, fortalece as inter relações entre as diferentes áreas do conhecimento.

Sendo assim, pretende-se com os objetivos apresentados anteriormente, desenvolver com base em uma metodologia conhecida, um algoritmo eficiente e preciso para a classificação monotônica de dados correlacionados.

1.3 Estrutura da tese

O tópico 2 desta tese apresenta uma revisão da literatura em problemas de classificação, com destaque para os problemas de classificação ordenada, principais metodologias e medidas de avaliação. No tópico 3 é apresentado o método proposto, sua compilação no formato de pacote, além da definição da metodologia de simulação que permitirá a comparação entre a metodologia proposta e a de referência, CPP-tri. Além dos parâmetros do estudo de simulação, é apresentado um conjunto de dados reais que será utilizado para validar os resultados do estudo simulado.

O tópico 4 é dividido em duas seções: a primeira apresenta um estudo de simulação no qual o método proposto é comparado com a metodologia CPP-tri sob igualdade de condições e apresentados os resultados desta comparação; a segunda

seção trata da aplicação das metodologias em estudos em um conjunto clássico de dados reais.

O tópico 5 trás as principais conclusões deste estudo, bem como propostas de continuidade.

2 REFERENCIAL TEÓRICO

2.1 Classificação

Em muitos estudos estatísticos, quer seja de natureza experimental ou observacional, é encontrado problemas que o objetivo principal é o de alocar um novo indivíduo[1] em uma, de duas ou mais classes conhecidas *a priori*, com base em um vetor de observações multivariada (FERREIRA, 2018). De outro modo, Requena (2018) esclarece que problemas dessa natureza são qualificados como *problemas de classificação*, e neles, deseja-se construir uma função que, baseada nas características desse indivíduo, o prediz em uma das classes ou categorias que ele pertence. O modo como, em geral, a comunidade estatística aborda tal problema, é através do desenvolvimento de metodologias de classificação, que relacionam uma variável qualitativa, comumente denominada como classe, com as variáveis ditas características ou atributos (JAMES et al., 2013).

Basgalupp (2010) define um atributo como uma variável observável e independente, que representa a característica ou propriedade apresentada pelo indivíduo. Por exemplo, a renda média familiar, idade e classe social, são atributos observados em investidores por um determinado banco. Conforme MORETTIN (2017), a natureza do atributo pode ser classificada em dois tipos: *qualitativo* ou *quantitativo*.

As possíveis realizações de um atributo *qualitativo* ou categórico, expressam uma qualidade ou característica do indivíduo pesquisado (MORETTIN, 2017). Por exemplo: o tipo sanguíneo de um paciente (A,B,AB,O) ou estado civil (casado, divorciado, solteiro). Conforme Frizzarini (2015), esse tipo de atributo pode representar uma característica que comporta-se de maneira ordenada, como o *grau de instrução* e seus possíveis resultados (fundamental, médio e superior), em alguns casos essa ordenação é representada numericamente, contudo, não é permitido que sejam realizadas operações aritméticas. MORETTIN (2017) define categoricamente os dois tipos de atributo qualitativo:

Qualitativo nominal: são os atributos cujos valores identificam nomes, rótulos ou categorias. Os valores de um atributo do tipo nominal não tem uma ordem natural e não devem ser utilizados em operações aritméticas, mesmo que sua represen-

[1] Embora alternativa, elemento, item, objeto sejam sinônimos na literatura, neste trabalho será adotado o uso do termo indivíduo.

tação seja numérica.

Qualitativo ordinal: O nível ordinal é composto por atributos cujos valores podem ser organizados em alguma ordem. Entretanto, a magnitude das diferenças entre as classificações não pode ser mensurada.

Os atributos *quantitativos* apresentam como possíveis realizações números resultantes de contagem ou mensuração. Segundo MORETTIN (2017), isso configura dois tipos de atributo quantitativo,

Quantitativo discreto: cujos possíveis valores formam um conjunto finito ou enumerável de números, e que resultam, frequentemente, de uma contagem, como por exemplo número de filhos $(0, 1, 2, \cdots)$;

Quantitativo contínuo: cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração, como por exemplo a temperatura corporal de um determinado indivíduo.

Assim como em Faceli et al. (2011), os conceitos abordados no problema de classificação serão esclarecidos com o auxílio de exemplo prático, mais precisamente, um conjunto de dados com informações do estado de saúde de alguns pacientes. O autor titula esse conjunto de dados como *hospital*, e as informações contidas nessa base estão organizadas na Tabela 1, e referem-se aos sintomas apresentados por cada paciente, assim como, os resultados de exames clínicos. Cada indivíduo ou paciente está arranjado em linha e os atributos em colunas, com informações de idade, temperatura corporal ($^{\circ}C$), presença de quadros específicos como febre, enjoos e distribuição de manchas na pele. Esses atributos também são chamados de covariáveis, critérios, variáveis preditoras ou também de variáveis independentes (BASGALUPP, 2010; REQUENA, 2018).

Tabela 1: Conjunto de dados *hospital* (adaptada de Faceli et al., 2011 e Basgalupp, 2010).

Paciente	Idade	Temp. ºC	Febre	Enjôo	Manchas	Dor	Diagnóstico
1	28	38	sim	sim	pequenas	sim	doente
2	49	38	não	não	grandes	não	saudável
3	21	37,6	sim	sim	pequenas	não	saudável
4	18	39,5	sim	não	grandes	sim	doente
5	34	38,4	sim	não	pequenas	sim	saudável
6	18	38,5	não	não	grandes	sim	doente

O atributo *Diagnóstico* no exemplo considerado (Tabela 1), corresponde ao que denomina-se como classe no problema de classificação. Nas áreas de estatística, e ciências da computação, especialmente em contextos de inteligência artificial e aprendizagem de máquinas, são encontrados diversas denominações para a classe, os autores Breiman (2001), Faceli et al. (2011), Silva, Barros e Costa (2016) e Requena (2018) apontam para termos como atributo alvo, categoria, variável resposta ou variável dependente.

A classe representa uma característica que pode ser influenciada ou talvez explicada por outros atributos. Na base de dados *hospital*, o estado de saúde de cada paciente é indicado pelo atributo (*Diagnóstico*), que pode apresentar resultados influenciados por sintomas de febre, enjoos e dores. Os rótulos das classes (*Doente* e *Saudável*) identificam categorias ou classes às quais os indivíduos pertencem (FACELI et al., 2011).

Formalmente, os dados observados em n indivíduos e relativos a m variáveis são representados em uma matriz $\boldsymbol{X}$ de dimensão $(n \times m)$

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nm} \end{pmatrix} \tag{1}$$

$$= \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_i^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{(1)} & \dots & \boldsymbol{x}_{(j)} & \dots & \boldsymbol{x}_{(p)} \end{pmatrix} \tag{2}$$

o vetor linha $\boldsymbol{x}_i^T$, representa o vetor m-dimensional de observações do i-ésimo indivíduo, e o vetor coluna $\boldsymbol{x}_{(j)}$ representa o vetor n-dimensional de observações correspondente a j-ésima variável, com $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, m$.

Existem várias definições diferentes para a matriz $\boldsymbol{X}$. O importante é que qualquer definição de $\boldsymbol{X}$ tenha a propriedade de que o vetor de medida $\boldsymbol{x}_i^T$ seja um indivíduo ou objeto que deseja-se classificar (VAPNIK, 1995).

Os indivíduos podem assumir representações distintas, e que dependem exclusivamente do contexto em que o problema de classificação será solucionado (THEODORIDIS, 2010; NETO et al., 2015; FACELI et al., 2011).

A atribuição de indivíduos (alternativas, elementos, observações ou objetos) à grupos homogêneos é um problema de grande interesse prático e de pesquisa (ZOPOUNIDIS; DOUMPOS, 2002). Nesse sentido, uma grande variedade de abordagens foi adotada para essa tarefa. As principais áreas de pesquisa podem ser identificadas: estatística, econometria, inteligência artificial e pesquisa operacional (FACELI et al., 2011).

De uma forma geral, estudos que envolvem o desenvolvimento de novas metodologias de classificação de dados, são realizados sob as perspectivas de produzir um classificador preciso ou descobrir a estrutura preditiva do problema (BREIMAN, 2001). No segundo caso, tenta-se entender quais variáveis ou interações de variáveis conduzem o fenômeno, ou seja, pretende-se fornecer caracterizações simples das condições que determinam quando um indivíduo/objeto está em uma classe e não em outra. Esses dois não são exclusivos, e na maioria das vezes, os objetivos serão tanto por uma previsão quanto um entendimento. Às vezes, um ou outro terá maior ênfase (BREIMAN et al., 1984; VAPNIK, 1995).

A classificação de dados possui aplicações em diversas áreas. Alguns trabalhos recentes envolvem detecção de mensagens de spam em *e-mails* baseada em características do cabeçalho e conteúdo da mensagem (SHARAFF; NAGWANI; DHADSE, 2016), predição do diagnóstico e identificação imunológica de pacientes com câncer de mama (SUZUKI et al., 2019), classificação de grãos de café baseado na cor que eles apresentam(OLIVEIRAL DIMAS SAMID LEME, 2016), análise de risco de crédito baseada em informações pessoais e financeiras do cliente (IMTIAZ; BRIMICOMBE, 2017), entre outros.

Yevseyeva (2007) enfatiza que as informações sobre as classes são fundamentais para o entendimento das especificidades de cada tipo de problema de classificação. Geralmente, quando existem métodos científicos, existem pequenas variações de teoria, metodologia e termos (GENTLE, 2009). Por esse motivo, esse trabalho preocupou-se em apresentar no decorrer do texto, termos que são comumente utilizados pela comunidade estatística, inteligência artificial e pesquisa operacional.

2.1.1 Classificação supervisionada

A classificação supervisionada ou problema de classificação, se dá quando as informações à respeito das classes são bem descritas e estabelecidas *a priori*, ou seja, o número de classes k e seus indivíduos são perfeitamente identificados antes de qualquer metodologia ser aplicada (FERREIRA, 2018). Nesse problema, concentra-se em alocar um novo indivíduo a uma das classes ou populações existentes, considerando uma função objetivo do vetor de observações multivariada x desse indivíduo (BOUVEYRON et al., 2019; KASSAMBARA, 2017).

Tabela 2: Exemplo de um conjunto de dados de empréstimo bancário.

cliente	renda	educação	registro.criminal	empréstimo
1	baixa	baixa	razoável	não
2	baixa	baixa	excelente	baixa
3	média	intermediária	excelente	intermediária
4	alta	baixa	excelente	alta
5	alta	intermediária	excelente	alta

Segundo Bouveyron et al. (2019), o termo "supervisionada" indica que as informações à respeito das classes são definidas antecipadamente, com base talvez na experiência e supervisão de um especialista, ou até mesmo por aplicação de metodologias que determinam quais grupos estão presentes nos dados (MICHAEL; CONSTANTIN, 2002; GENTLE, 2009). Em outras palavras, em um problema de classificação, o analista sabe antecipadamente como devem ser os resultados da análise. Por exemplo, suponha que um banco queira basear sua política de empréstimos em várias características apresentadas por seus clientes, tais como, renda, nível de escolaridade e antecedentes criminais. Se um cliente recebe um empréstimo, ele pode pertencer a uma dessas três classes que descrevem a quantidade em dinheiro concedido pelo banco: baixa, intermediária e alta. Portanto, junto com a opção "sem empréstimo", são estabelecidas quatro classes (Tabela 2). Suponha ainda que o banco deseje basear sua política de empréstimos em classes definidas a partir de um número de decisões sobre dignidade de crédito no passado, configurando dessa maneira um problema de classificação supervisionada.

Nesse sentido, Bouveyron et al. (2019), Kassambara (2017) ressaltam que para esse contexto a classificação dos indivíduos é induzida pelas informações já conhecidas à respeito das classes. Por exemplo, no campo da ecologia tem-se o interesse na observação remota de vida selvagem. Com esse propósito, os pesquisadores Tabak et al. (2018), Willi et al. (2019), instalaram em pontos estratégicos armadilhas fotográficas. Essas armadilhas são compostas por câmeras que são ativadas por movimento, e o registro de milhões de imagens dos amimais silvestres são capturadas por elas (Figura1). A classificação dessas imagens foi realizada de forma supervisionada, já que todas as características que descrevem as classes de animais silvestres são conhecidas e utilizadas pelo o procedimento de classificação.

Figura 1: Exemplos de imagens de armadilhas fotográficas utilizadas em estudos com animais silvestres.

Classificação supervisionada é um dos problemas mais estudados na área de aprendizagem de máquina. Atualmente existem excelentes métodos disponíveis, que vão desde os mais simples, como a regressão logística, até os mais sofisticados, como as Florestas Aleatórias (*Random Forests*) e as Máquinas de Suporte de Vetores (*Support Vector Machines*)[2].

2.1.2 Classificação não supervisionada

O problema de classificação não supervisionada surge em situações nas quais as classes não são definidas *a priori*, nenhuma informação quanto à quantidade k de classes é conhecida, muito menos, como os indivíduos estão agrupados segundo suas características. Esse tipo de problema, é denotado por muitas áreas como problema de agrupamento (*clustering*) (GENTLE, 2009; FACELI et al., 2011; KASSAMBARA, 2017).

Os autores Ferreira (2018), Everitt, Landau e Leese (2009) enfatizam que, diferentemente do problema de classificação supervisionada, em um problema de agrupamento a classificação dos indivíduos não é guiada por informações *a priori* de quais variáveis ou amostras pertencem a quais classes, população ou grupo. A classificação não supervisionada agrupa indivíduos com base apenas nas informações encontradas nos dados que descrevem os objetos e seus relacionamentos.

O intuito da classificação não supervisionada é que os indivíduos dentro de um grupo sejam semelhantes (ou relacionados) entre si e diferentes (ou não relacionados) aos objetos de outros grupos. Quanto maior a semelhança (ou homogeneidade) dentro de um grupo e maior a diferença entre os grupos, melhor ou mais distinto o agrupamento (GENTLE, 2009; AGGARWAL; REDDY, 2013).

Como evidenciado no trabalho de Kassambara (2017), a identificação de grupos ou de observações semelhantes em um conjunto de dados é uma etapa impor-

[2]conhecido como SVM

tante para entender os dados e entender os fenômenos representados por eles.

Considera-se indivíduos semelhantes aqueles que são alocados em uma mesma classe e, por isso, aqueles que pertencem a diferentes classes são considerados dissimilares (FERREIRA, 2018). A semelhança entre os indivíduos é quantificada por meio de medidas de proximidade, que engloba tanto as medidas de similaridades quanto as de dissimilaridade.

O trabalho dos autores (TAN et al., 2018), apresentam os conceitos básicos e uma visão geral dos algoritmos mais influentes relatados no campo da classificação não supervisionada, destacando suas limitações ou desvantagens.

Aggarwal e Reddy (2013) afirma que por entendimento ou utilidade, a classificação não supervisionada tem desempenhado um papel importante em uma ampla variedade de campos: medicina, psicologia e outras ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informações, aprendizado de máquina e mineração de dados

2.1.3 Classificação binária

A classificação binária é considerada o problema de classificação mais simples dentre os casos existentes. As instâncias em conjuntos de dados para classificação binária apresentam apenas uma variável resposta, que possui somente dois níveis de classificação. Estes valores de modo geral são positivos e negativo, mas podem ser interpretados como verdadeiros e falsos, codificados usualmente como 1 e , ou combinações destes dois valores. Um exemplo clássico é apresentado em Herrera et al. (2016), tratando da classificação de mensagens de e-mail como *spam*.

Um classificador binário tem por objetivo encontrar limiares capazes de efetuar a separação das instâncias em dois grupos, pertencentes às classes denominadas positivas ou negativas, verdadeiras ou falsas, etc. Algumas das aplicações clássicas desta metodologia tratam da concessão de crédito, avaliação médica, reconhecimento de padrões, dentre outras aplicações. A Figura2 apresenta um desenho sistemático desta metodologia de classificação.

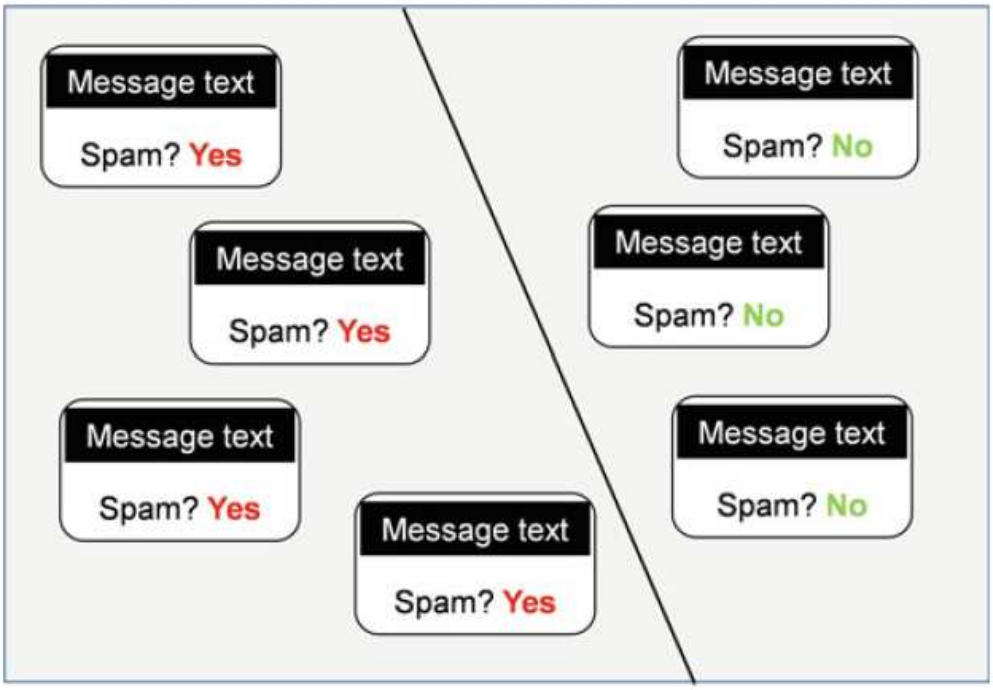

Figura 2: Exemplo de classificação binária aplicada à detecção de *spam*

2.1.4 Classificação multiclasses

Na abordagem de classificação multiclasses, os conjunto de dados também apresentam apenas uma variável resposta, como na classificação binária. Entretanto, diferentemente da classificação binária, esta variável pode conter um conjunto de possíveis valores pré-definidos, de modo geral, um conjunto discreto, sendo que o valor de cada possível variável pode representar valores específicos para a aplicação em contexto. Além de discreto, é necessário que o conjunto seja finito, pois caso contrário se trataria de um problema mais apropriado para metodologias de regressão em detrimento à classificação. De modo geral, considera-se a abordagem multiclasses uma generalização da abordagem de classificação binária, na qual retira-se a restrição da presença de apenas duas classes possíveis de resposta, podendo esta assumir, conforme supracitado, valores contidos em um conjunto finito e discreto de valores. A Figura3 obtida de Herrera et al. (2016) apresenta um diagrama referente à aplicação do método multiclasses ao clássico problema de classificação de espécies de iris.

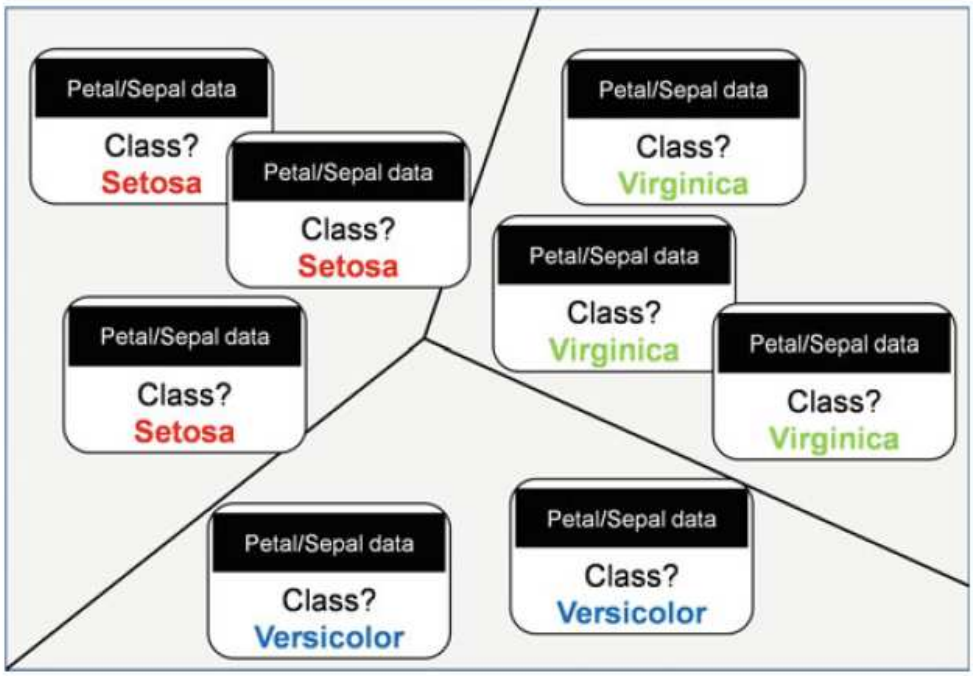

Figura 3: Exemplo de classificação multiclasses aplicada ao problema de classificação de espécies de iris

2.1.5 Classificação nominal

Quando as classes não são ordenadas, o problema de classificação consiste em um problema de caráter nominal. Neste caso, os objetos pertencentes a diferentes grupos possuem características diferentes, sem que seja possível estabelecer qualquer tipo de relação de preferência entre elas. Um estudo pioneiro relacionado a classificação nominal foi desenvolvido por Fisher (1936). Nesse trabalho, o autor utilizou análise discriminante na classificação da íris.

2.1.6 Classificação ordinal

Na classificação ordinal, as classes apresentam uma ordenação natural. Neste tipo de problema, cada entrada de um vetor de dados é atribuída a um conjunto discreto de categorias ordenadas. Este tipo de problema é relacionado, de modo geral, a informações que apresentam uma estrutura de ordenação naturalmente definida, como por exemplo, risco de investimentos em um ativo, que pode ser classificado como "baixo", "médio"e "alto", classificação de produtos por sua qualidade, como por exemplo variedades de arroz, que podem ser definidas por sua qualidade em função da quantidade de grãos defeituosos, sendo 1 o que apresenta menos grãos com ca-

racterísticas negativas enquanto o tipo 5 apresenta maior quantidade de elementos indesejados, dentre diversos outros exemplos.

Cano, Luengo e García (2018) reforçam que a ordenação das classes pode ser explorada para construir mais classificadores precisos em aplicações que envolvem preferências, escolha social, tomada de decisão com vários critérios ou decisão sob risco e incerteza. Por exemplo, considere uma fábrica, um funcionário pode ser avaliado como "excelente", "bom"ou "ruim"ou um risco de crédito pode ser classificado como "AAA", "AA", "A"ou "A-".

Gutiérrez e García (2016) apresentam as principais abordagens para metodologias de classificação monotônica, bem como, exemplos de aplicação das referidas metodologias. Quando dentro do problema de classificação ordinal se busca inserir restrições que culminam em relações de monotonicidade, se diz que o problema é de classificação monotônica.

2.1.6.1 Classificação monotônica

Em muitas aplicações de análise de dados, é razoável supor que a variável resposta esteja aumentando (ou diminuindo) em uma ou mais variáveis ou recursos. Tais relações entre resposta e atributo são chamadas monótonas.

Nesse tipo de problema, explora-se as propriedades de ordenação dos vetores de atributos, usando o conhecimento disponível em termos de dominância. Conforme (CANO; LUENGO; GARCíA, 2018), uma amostra domina a outra quando cada coordenada da primeira não é menor que a respectiva coordenada da última.

As restrições de monotonicidade exigem que o rótulo de classe atribuído a um indivíduo seja maior ou igual ao rótulo de classe atribuído aos indivíduos que ele domina (CANO et al., 2018). Como exemplo, considere uma restrição de monotonicidade relacionada a um atributo de entrada e à classe de destino. Nesse caso, uma amostra no conjunto de dados com um valor mais alto do atributo de entrada não deve ser associada a um valor de classe mais baixo, desde que os outros atributos da amostra sejam corrigidos. Essas restrições de monotonicidade podem ser diretas (como o exemplo apresentado anteriormente) ou inversas (se o valor do atributo diminuir, o valor da classe não deve aumentar).

A classificação com restrições de monotonicidade, também conhecida como

classificação monotônica (GUTIÉRREZ; GARCÍA, 2016), é um problema de classificação ordinal no qual busca-se impor uma restrição monotônica: a presença de um valor com maior módulo em um conjunto de dados, fixados os demais, não apresenta impacto negativo em sua atribuição de classe. Neste tipo de problema, a monotonicidade presente entre as variáveis dependentes e independentes, são comumente utilizadas como conhecimento prévio no momento da classificação dos dados em estudo (KOTŁOWSKI; SŁOWIńSKI, 2013). Podemos considerar as avaliações de alunos em uma universidade. Os alunos são avaliados com uma classificação entre 0 e 10. Como um exemplo descritivo, considere três alunos (Estudantes A, B e C), cada um com 22 avaliações e uma nota final. Considere também que todos os atributos (22 avaliações) tem uma suposição monotônica direta em relação aos possíveis resultados da classe, nesse caso, a nota final representada em negrito:

- Estudante A: $5, 5, 5, 5, 7, 6, 5, 5, 5, 5, 5, 5, 6, 5, 5, 6, 6, 6, 5, 5, 5, 5, \mathbf{4}$
- Estudante B: $3, 5, 3, 4, 7, 3, 3, 5, 3, 3, 3, 3, 6, 3, 3, 4, 3, 6, 4, 3, 5, 3, \mathbf{5}$
- Estudante C: $2, 2, 1, 2, 1, 2, 2, 3, 2, 2, 1, 2, 3, 2, 2, 3, 3, 2, 2, 1, 2, 3, \mathbf{2}$

Como pode ser observado, há uma violação monotônica envolvendo duas amostras (Estudante A e B), onde o Estudante B, que tem notas de avaliação piores ou iguais às do Estudante A, apresenta uma nota final mais alta. Por outro lado, não há violações monotônicas ao considerar o aluno C em relação aos alunos A e B.

Agora, definimos formalmente um conjunto de dados de classificação com rótulos ordinais com restrições de monotonicidade. Vamos supor que os indivíduos sejam descritos usando um total de m atributos com domínios ordenados, $\boldsymbol{x}_i^T \in \boldsymbol{R}^m$, e os rótulos das classes, y_i, de um conjunto finito $\mathbf{C}$ de rótulos ordenados, $y_i \in \mathcal{Y} = \{1, \cdots, \boldsymbol{C}\}$ Como discutido anteriormente, uma relação de dominância $\preceq$ é definido da seguinte forma:

$$\boldsymbol{x}_i^T \preceq \boldsymbol{x}_i^{T'} \iff x_{im} \geq x_{i,m'}, \forall\, j = 1, \cdots m, \forall\, j = 1, \cdots n. \tag{3}$$

sendo x_{im} e $x'_{i,m'}$ as m-ésimas coordenadas dos indivíduos $\boldsymbol{x}_i^T$ e $\boldsymbol{x}_i^{T'}$, respectivamente. Em outras palavras, $\boldsymbol{x}_i^T$ domina $\boldsymbol{x}_i^{T'}$ se cada atributo $\boldsymbol{x}_i^T$ não ter resultado menor que os respectivo atributo medido em $\boldsymbol{x}_i^{T'}$.

Segundo Cano, Luengo e García (2018), os dados ordinais monotônicos podem ser definidos da seguinte maneira. Seja $\boldsymbol{X}$ um conjunto de dados com p variáveis ou

atributos ordinais $A_1, \cdots, A_p$ e uma variável classificadora $Y = \{C_1, C_2, \cdots, C_k\}$ com r possíveis valores ordinais. O conjunto de dados consiste de n indivíduos x_i.

Cano et al. (2018) afirma que a ordenação entre as classes podem ser explorada para construir métodos mais acurados nos domínios de aplicações que envolvem preferências, como escolha social, tomada de decisão com vários critérios ou decisão sob risco e incerteza. Em muitos cenários do mundo real, como previsão de falência, precificação de opções ou diagnóstico médico, em que os modelos de classificação a serem aprendidos precisam cumprir restrições de monotonia. Por exemplo, é racional supor que uma maior taxa de juros da empresa nunca deve resultar em um nível mais baixo de risco de falência. Consequentemente, existe um interesse crescente da comunidade de pesquisa em estatística em relação a modelos preditivos monotônicos.

O trabalho de Cano et al. (2018) apresenta uma revisão aprofundada das métricas utilizadas na avaliação de modelos de classificação monotônica. Neste trabalho, observa-se que a metodologia predominante para a avaliação da qualidade dos resultados obtidos é a acurácia, sendo esta medida presente em 47% dos trabalhos estudados.

Michael e Constantin (2002), descrevem a relação entre as classes e os diferentes tipos de problema de classificação, e um pouco de suas especifidades com um fluxograma (Figura4).

A Figura5, adaptada de Cano et al. (2018), indica o crescente interesse na metodologia de classificação monotônica na comunidade acadêmica nos últimos 10 anos, indicando que cada vez mais metodologias que resolvam problemas de dados monotonicamente relacionados vem se tornando relevantes cientificamente.

Na seção seguinte, será apresentado sucintamente o método de composição probabilística de preferências, com o intuito de melhor descrever o método no qual a metodologia proposta nessa tese baseou-se, o CPP-tri.

2.2 O método CPP

O método de Composição Probabilística de Preferências (CPP)[3] desenvolvido inicialmente pelos autores Sant'anna e Sant'anna (2001) e ampliado, recentemente por Sant'anna (2015). Baseia-se na teoria de probabilidades conjuntas em apoio à

[3]do inglês, *Probabilistic composition of preferences*

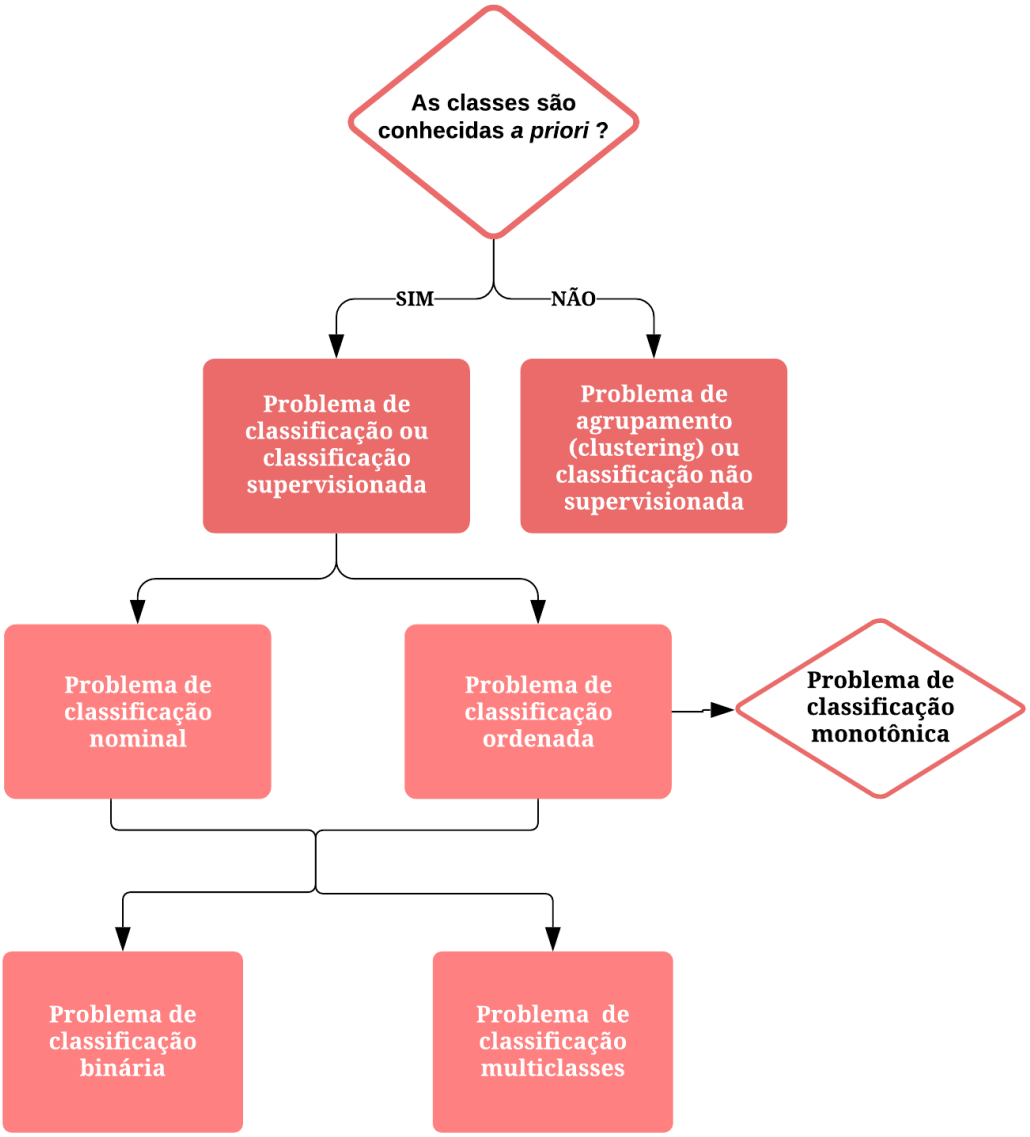

Figura 4: Problemas de classificação. (Adaptada de Michael e Constantin (2002))

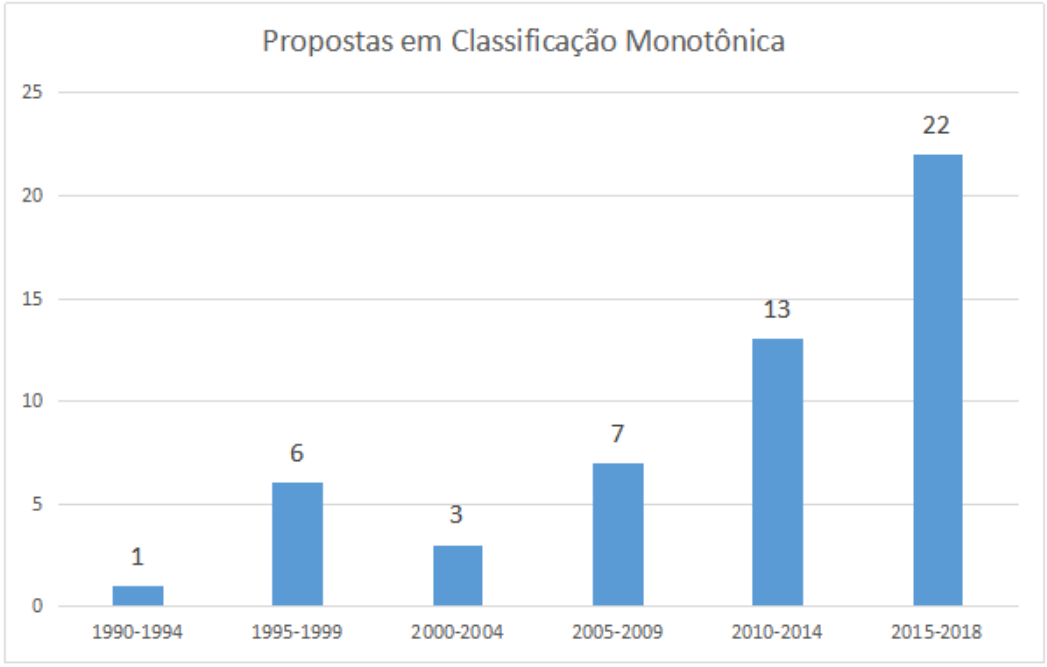

Figura 5: Propostas de trabalhos envolvendo classificação monotônica (adaptado de Cano et al. (2018))

problemas de decisão , especialmente na escolha, ordenação e classificação de alternativas (e.g. indivíduos) avaliadas sob múltiplas características ou critérios.

Os autores Roy e Skalka (1984) classificam situações de decisão em quatro problemáticas:

Ordenação *(ranking)* na qual visa construir uma lista ordenada das alternativas consideradas, das melhores para as piores. O autor denota essa problemática como (P.γ).

Escolha *(choice)* na qual visa identificar a melhor alternativa dentre as consideradas, ou um conjunto limitado das melhores alternativas. O autor denota essa problemática como (P.α).

Descrição *(description)* que consiste em identificar e descrever as principais características significativas na distinção das alternativas. O autor denota essa problemática como (P.δ).

Classificação *(classification/sorting)* cujo o intuito é alocar as alternativas disponíveis, em classes estabelecidas *a priori* que preservam uma relação de ordem quanto à preferência. O autor denota essa problemática como (P.β).

O CPP fundamenta-se na ideia de que ao se aplicar um critério de preferência para tomar uma decisão, existe um componente subjetivo que torna imprecisa a me-

dida de preferência, mesmo que baseada na observação de uma variável determinada objetivamente com toda precisão (SANT'ANNA, 2013c).

A imprecisão na qual refere-se Sant'Anna (2013b), impõe que as medidas de preferência segundo cada critério sejam examinadas como variáveis aleatórias. Possibilitando por exemplo, que as preferências possam ser medidas em termos de probabilidades, de ser a melhor opção em uma problemática de escolha, ou de ser melhor ou pior que os perfis pré-estabelecidos como referência das classes no problema de classificação. Cada alternativa é identificada por um vetor, e entende-se como preferência cada coordenada desse vetor, segundo um critério diferente (SANT'ANNA, 2015)

A escolha de distribuição de probabilidades que melhor se ajusta aos dados pode advir de informações a priori, de aplicações bem-sucedidas a problemas similares, da avaliação do ajuste dos dados a distribuições conhecidas ou, na ausência dessas informações, arbitrada ao contexto. De maneira geral, aplicações do método com outras distribuições no lugar da normal podem ser empregados sem mudanças substanciais no procedimento, no entanto, não é esperado grandes divergências nos resultados (SANT'ANNA; COSTA; PEREIRA, 2012). Esse comportamento foi constatado no estudo de caso multivariado, indicando que a abordagem com distribuições contínuas apresentou melhores resultados (GAVIÃO et al., 2019).

Uma revisão sumária da literatura indica que o CPP tem modelado problemas com distribuições de probabilidades normais (SANT'ANNA, 2013a; SANT'ANNA; COSTA; PEREIRA, 2015), uniformes (SANT'ANNA; CONDE, 2011; SANT'ANNA; FERREIRA; DUARTE, 2012), triangulares (TREINTA et al., 2014), de Pareto (CAILLAUX et al., 2011; SANT'ANNA; MELLO, 2012) e Beta (SANTANNA et al., 2015; SANT'ANNA, 2015).

Gavião et al. (2019), enfatiza que a natureza probabilística do CPP é uma característica muito importante para o tratamento de dados imprecisos. Segundo Sant'Anna, Faria e Costa (2013), essa imprecisão pode decorrer de diferentes processos que envolvem avaliações de especialistas, de medidas de desempenho imprecisas, de processos com sistemas métricos imperfeitos, dentre outros que envolvem avaliações humanas em situação de incerteza.

Recentemente foram propostas algumas variações da metodologia, Martins

(2015) propôs em sua tese uma metodologia baseada na CPP, denominado com CPP-Hibrido. Esse instrumento, considera aspectos relativos aos erros humanos em um ambiente de risco, e suas possíveis contribuições às falhas em instalações de alta complexidade. O foco principal dessa metodologia é investigar a confiabilidade humana e a identificação de parâmetros a serem controlados para reduzir ou detectar a condição de falha. Recentemente, Gavião et al. (2019) propuseram uma abordagem empírica ao método de Composição Probabilística de Preferências (CPP). O uso de probabilidades empíricas, sem a necessidade de conhecer ou assumir uma função de probabilidade conjunta que origina as preferências, pode contribuir para lidar com determinados tipos de problema com escalas ordinais de preferência.

A Tabela 2 apresenta estudos que aplicaram a metodologia CPP em problemas de classificação ordenada. Nesse caso em específico, a metodologia é nomeada como CPP-tri e apresentada na seção 2.3.

A composição probabilística de preferências permite combinar as avaliações probabilísticas por diferentes perspectivas (SANT'ANNA; SANT'ANNA, 2001). Em problemas de classificação, essas abordagens de composição substituem a necessidade de atribuição de pesos aos critérios, evitando assim, que a classificação das alternativas seja realizada de maneira subjetiva.

2.3 CPP-tri

O método CPP-tri proposto por (SANT'ANNA, 2015), tem por objetivo alocar cada alternativa (indivíduo) em uma de um conjunto de classes ordenadas. Por se tratar de um método de classificação supervisionada, as classes pertencentes ao problema que o método será aplicado devem ser estabelecidas *a priori*.

Neste método, as classes são identificadas por vetores multivariados, nomeados como perfis de referência ou perfis representativos. Para levar em conta as relações entre os m critérios,os perfis representativos podem ser construídos de várias formas e em qualquer número. Por exemplo, cada classe pode ser representada por um limite superior e um limite inferior (SANT'ANNA; COSTA; PEREIRA, 2015). A Tabela 3 apresenta uma relação de problemas abordados por meio da metodologia CPP-TRI.

Segundo (SANT'ANNA, 2013c), as avaliações que compõem os perfis de refe-

Tabela 3: Publicações do método CPP-tri

AUTORIA	DESCRIÇÃO
Sant'Anna et al. (2013)	Aplicação da Composição Probabilística e do Método das k-médias à classificação de municípios quanto à oferta de creches
Faria et al. (2013)	Comparação entre resultados da Composição Probabilística de Preferências e da formação de agrupamentos pelo Método das k-médias
Sant'Anna et al. (2015)	CPP-TRI: a sorting method based on the probabilistic composition of preferences
Ribeiro et al. (2015)	Probabilistic Preferences Composition in the Classification of Apparel Retail Stores
Sant'Anna (2015)	Probabilistic Human Development Indices
Sant'Anna (2015)	Probabilities in the Problem of Classification
Silva et al. (2016)	Abordagem híbrida multicritério – mineração de dados aplicada a classificação de unidades da federação com base na população economicamente ocupada
Silva e Costa (2017)	Análise experimental do método de Composição Probabilística de Preferências em seu modo Tricotômico (CPP-tri) na classificação ordenadade vinhos da região do Minho
Gavião et al. (2017)	Uma nova abordagem aplicada ao conceito moneyball com apoio da Composição Probabilística de Preferências
Sant'Anna et al. (2017)	A probabilistic approach to the inequality adjustment of the human development index
Gavião et al. (2017)	Aplicação da Composição Probabilística de Preferências e do Índice de Gini à escolha de jogadores da Liga Inglesa de Futebol.
Sant'Anna (2017)	Aplicação da CPP-tri à classificação dos países pelos critérios do IDH
Monte e Silva(2018)	Aplicação da Composição Probabilística de Preferências Tricotômica (CPP-tri) para classificação do nível de maturidade na gestão de processos de negócios

rência são obtidas a partir da amostra de indivíduos a ser classificada. Por exemplo, perfis centrais seriam formados por coordenadas dadas pelos quantis, ou valor médio para cada critério ou variável considerada no estudo. Sant'Anna, Costa e Pereira (2015) ressalta que os perfis são determinados sob restrição de monotonicidade e disponibiliza um algoritmo para a ordenação desses perfis, em outras palavras, se um perfil representativo da primeira classe apresentar uma avaliação de acordo com um determinado critério, maior que o de um perfil em outra classe, as avaliações de todos os perfis da primeira classe devem ser superiores aos da segunda classe.

Sant'Anna (2013c) afirma que, adotar a mesma quantidade de perfis em cada classe, ou seja, $n(i) = n$, facilita a comparação das distâncias probabilísticas entre classes distintas. Essa comparação compreende uma etapa do método, apresentada na seção 2.3.1. Caso contrário, são adicionados perfis com valores iguais às médias aritméticas dos valores dos perfis iniciais.

Sant'Anna (2015) enfatiza que, nesta etapa em específico, deve-se evitar critérios que não apresentam um desempenho significativo na discriminação das alternativas.

Quanto à escolha da distribuição de probabilidades das pertubações, Sant'Anna (2013b) reforça que a distribuição normal é a mais utilizada para representar erros de medida. Na falta de informação precisa quanto ao parâmetro de escala da distribuição normal, estima-se esse parâmetro a partir da variância observada nas classes, segundo cada critério (SANT'ANNA, 2015).

A aplicação do método CPP-tri é realizado mediante as informações de quantidade r de classes, $n(i)$ perfis e escolha de uma distribuição que seja identificada ou mesmo assumida como característica das pertubações.

Formalmente, considera-se o problema de classificar uma alternativa A segundo o l-ésimo critério, com l variando de $1, \cdots, m$ e m denotando o número de critérios, em uma dentre k classes ordenadas $C = \{c_1, \cdots, c_r\}$ de r classes ordenadas, cada uma identificada por um certo número n de perfis de referência, constituído por avaliações de alternativas previamente construídas.Denote-se por C_{ijl} a avaliação do l-ésimo critério que aparece no j-ésimo perfil da i-ésima classe. Os perfis são definidos de modo que as classes estão efetivamente ordenadas em ordem crescente, com a restrição de que os perfis sejam representados por avaliações em todos crité-

rios menores ou iguais às dos perfis das classes acima dela. A aplicação do método CPP-tri envolve três etapas:

2.3.1 Cálculo das probabilidades de preferência

As medidas exatas a_l e C_{ijl} são usadas como médias de distribuições de variáveis aleatórias X_l e Y_{ijl}.

$$A_{il}^{-} = P(\boldsymbol{X}_l < \boldsymbol{Y}_{ijl}) \tag{4}$$

$$A_{il}^{+} = P(\boldsymbol{X}_l > \boldsymbol{Y}_{ijl}) \tag{5}$$

2.3.2 Cálculo das probabilidades globais

A_{il}^{+} e A_{il}^{-} representam, respectivamente, a probabilidade do indivíduo A apresentar desempenho respectivamente acima e abaixo dos valores informados para o critério l nos perfis da classe i. Depois que essas probabilidades locais são conhecidas, é possível calcular as probabilidades conjuntas globais de um objeto estar acima ou abaixo dos perfis de cada classe, levando em conta todo o conjunto de critérios, utiliza-se a composição probabilística propostas por Sant'anna e Sant'anna (2001).

Por independência, as probabilidades de estar acima e abaixo de todos os perfis que identificam as classe são encontradas pelo produto das probabilidades de estar acima ou abaixo de cada perfil,

$$A_{i}^{-} = \prod_{l=1}^{m} P(\boldsymbol{X}_l < \boldsymbol{Y}_{ijl}) \tag{6}$$

$$A_{i}^{+} = \prod_{l=1}^{m} P(\boldsymbol{X}_l > \boldsymbol{Y}_{ijl}) \tag{7}$$

2.3.3 Alocação nas classes mais próximas

Sant'Anna (2013b) estabelece como requisito para a aplicação do método que os perfis estejam ordenados, isto é, para todos os critérios, nenhuma avaliação do perfil de uma classe superior pode ser menor que a avaliação do perfil de uma classe inferior no mesmo critério; e pelo menos um critério do perfil da classe superior deve ser maior que a avaliação do perfil da classe inferior

Depois de calculadas as probabilidades dos indivíduos estarem acima ou abaixo dos perfis de uma classe, é possível calcular a diferença entre essas probabilidades, que é aqui chamada de distância probabilística α. Essas distâncias são calculadas para todas as classes e o indivíduo é alocado para a mais próxima, ou seja, para a classe para a qual o valor absoluto de α é o mais baixo.

$$\alpha = |A_i^+ - A_i^-| \tag{8}$$

No entanto, as premissas dos autores consideram a suposição da independência das variáveis, porém, em várias áreas do conhecimento existem situações em que as variáveis apresentam dependência e faz-se necessário a resolução de problemas de classificação. Tal situação limita a amplitude de ação da metodologia proposta pelos autores, e abre espaço para a realização de estudos que visam melhorar a classificação a partir da análise de parâmetros estatísticos tais como a correlação, que permite a aplicação da metodologia em situações em que as variáveis não supõem independência. Dentro deste contexto é onde o presente trabalho encontra-se alocado.

2.4 Distribuição Normal Multivariada

A distribuição normal multivariada é uma generalização para várias dimensões da densidade normal univariada. No campo de estudos da análise multivariada, a distribuição normal para $p \geq 2$ dimensões desempenha um papel muito importante, já que esta distribuição representa uma aproximação adequada de distribuições populacionais e dados experimentais, além de ser utilizada em várias áreas como engenharia, psicologia e economia, e de servir para descrever qualquer conjunto de variáveis aleatórias de valores reais correlacionados.

Definição 2.4.1 (Distribuição normal multivariada)**.** *Sejam $X_1, X_2, \ldots, X_p$ variáveis aleatórias contínuas independentes tal que $X_i \sim N(\mu_i, \sigma_i^2)$, para $i = 1, 2, \ldots, p$. Então $\boldsymbol{X} = [X_1, X_2, \ldots, X_p]$ distribuição normal multivariada se sua função densidade conjunta é dada por:*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{9}$$

em que $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]' \in \boldsymbol{R}^p$ e $\boldsymbol{\Sigma}$ é uma matriz positiva definida dada por:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \ldots & \sigma_p^2 \end{bmatrix}. \tag{10}$$

2.5 Boston Housing Price

O conjunto de dados *Boston Housing Price*, foi compilado no trabalho *Hedonic Housing Prices and the Demand for Clean Air* no final dos anos 70, pelos economistas David Harrison e Daniel Rubenfeld (HERRISON; RUBINFELD, 1976). Os autores investigaram a disposição de pagar por melhorias na qualidade do ar, usando dados para o mercado imobiliário de *Boston, Massachusetts, EUA*. Esse estudo é instigado por conclusões documentadas em trabalhos semelhantes, que discutem os danos que altas concentrações de poluentes atmosféricos impõem à saúde humana, à vegetação, e a vários materiais.

Herrison e Rubinfeld (1976) obtiveram os dados de Área Estatística Metropolitana Padrão de Boston, do inglês ***S**tandard **M**etropolitan **S**tatistical **A**rea*. Era assim, na década de 70, como o escritório americano de gestão e orçamento [4] definia formalmente uma cidade central com uma grande população e sua região circundante, que pode incluir vários municípios adjacentes, ligados por fatores sociais e econômicos [5]

Esse conjunto de dados possui características que são relevantes para o exame

[4] Office of Management and Budget (OMB).

[5] Atualmente, denomina-se Área Estatística Metropolitana, do inglês ***M**etropolitan **S**tatistical **A**rea* - MSA.

de diversas metodologias (GILLEY; PACE, 1996). Em especial , como o estudo dessa tese se trata de uma metodologia para solucionar problemas de classificação monotônica em dados correlacionados, a normalidade apresentada pelos dados, a quantidade de variáveis mensuradas, o comportamento da relação entre as covariáveis e o desbalanceamento entre as classes, foram critérios decisivos para a escolha dessa base de dados.

Após a compilação original desse conjunto em (HERRISON; RUBINFELD, 1976), outros pesquisadores utilizaram esses dados por diferentes perspectivas, derivadas em sua maioria de disciplinas científicas como aprendizado de máquinas, economia, estatística e mineração de dados. Por exemplo, Belsley, Kuh e Welsch. (1980), utilizaram esses dados para examinar metodologias de identificação de dados influentes e fontes de colinearidade. Posteriormente, os pesquisadores Krasker, Kuh e Welsch (1983), Shankar e Carson (1988), examinaram estimativas robustas na presença de heterocedasticidade. Os autores Breiman e Friedman (1985), discutiram uma metodologia para estimar transformações ideais para regressão múltipla e correlação. Pace (1993), utilizou essa base na investigação de estimativas não paramétricas. Lange e Ryan (1989), descreveram um método para averiguação da adequação do modelo e, em particular, a suposição distributiva sobre os efeitos aleatórios. Breiman et al. (1984) aplicou metodologias de classificação e árvore de regressão. Ainda nesse contexto, recentemente os autores Shahhosseini, Hu e Pham (2019), Cano, Luengo e García (2018), Cano et al. (2018), consideraram a base *Boston Housing Price* na investigação de algoritmos de filtragem de ruído em um estágio de pré-processamento com um duplo objetivo: aumentar o índice de monotonicidade dos modelos e a precisão das previsões para diferentes classificadores monotônicos.

3 MATERIAL E MÉTODOS

Nesta seção, será apresentada a metodologia proposta, que difere essencialmente da metodologia CPP-tri ao levar em consideração a existência de correlações entre os atributos. Os experimentos realizados basearam-se na ideia de construir um ambiente com diferentes cenários de correlação entre os atributos e contrastar as acurácias das duas metodologias.

3.1 Dados reais

A metodologia proposta foi aplicada em um conjunto de dados reais, disponível gratuitamente para *download*, e intitulado como *Boston Housing Price* nos repositórios de base dados UCI (DUA; GRAFF, 2017) e GitHub (GITHUB, 2019). O repositório de aprendizagem de máquina da Universidade da Califórnia, Irvine *(UCI Machine Learning Repository)*, é um conjunto de bancos de dados de acesso livre e regularmente visitado por pesquisadores de áreas científicas, tais como, Estatística, Inteligência artificial, Mineração de dados, Apoio à decisão multicritério e outras. A UCI, fornece bases de dados artificiais e experimentais que viabilizam a análise, desenvolvimento e validação de metodologias pertencentes à diversas áreas cientificas.

O conjunto de dados *Boston Housing Price* é pequeno em tamanho, com 506 indivíduos. O conjunto de dados original apresenta um total de 14 variáveis sendo que, para este trabalho, foi selecionado um estrato de 7 destas. Uma breve descrição dessa base é fornecida a seguir. Herrison e Rubinfeld (1976) apresentam uma descrição mais detalhada.

MEDV Define o valor médio das casas ocupadas pelos proprietários em termos de milhares de dólares $ 10000;

RM A variável RM representa o espaço e, em certo sentido, quantidade de habitações. Deve estar positivamente relacionado ao valor da habitação;

LSTAT Proporção da população de nível socioeconômico baixo;

CRIM A taxa de criminalidade por cidade. Como o *CRIM* avalia a ameaça ao bem-

estar que as famílias percebem em vários bairros vizinhos da área metropolitana de Boston (assumindo que as taxas de crime são geralmente proporcionais às percepções de perigo das pessoas), isso deve ter uma relação negativa com os valores da habitação;

PTRATIO Relação aluno-professor por distrito escolar da cidade. Mede os benefícios do setor público em cada cidade. A relação entre a razão aluno-professor e a qualidade da escola não é totalmente clara, embora uma proporção baixa deva implicar que o aluno receba mais atenção individual. esperamos que o *PTRATIO* esteja negativamente relacionado aos valores da habitação ;

DIS Distâncias ponderadas para cinco centros de emprego na região de Boston. De acordo com as teorias tradicionais dos gradientes de aluguel de terrenos urbanos, os valores das moradias devem ser mais altos perto dos centros de emprego. Espera-se que a variável *DIS* esteja relacionada negativamente com os valores de habitação;

NOX Concentrações de óxido de nitrogênio em *pphm* (concentração média anual em partes por cem milhões).

GROUPS Variável gerada pelos autores para atuar como grupo de classificação. Para garantir que a variável gerada seja categórica ordinal, a variável **MEDV**, variável regressora no conjunto de dados original, foi dividida em quatro grupos em seus respectivos quartis, sendo que cada quartil gerou um grupo ordenado, culminando num total de 4 categorias ordenadas para os dados.

3.2 Método proposto

Denomina-se o método aqui proposto como CPP-Cor, seu nome faz referência a duas características importantes para o seu desenvolvimento, a metodologia em que ele se baseia (CPP), e a forma com qual as interações entre as variáveis estão sendo consideradas. O mesmo, com o intuito de fazer alusão a duas características importantes, apresenta finalidade similar ao método CPP-tri, no entanto, a classificação realizada pelo método CPP-cor leva em consideração a correlação entre as características observadas nos indivíduos. Tornando-o relevante a contextos

de elevada incerteza quanto a correlação dos dados, pois, essas relações entre as variáveis são desconsideradas a partir da pressuposição de independência, estabelecida pelo método CPP-tri. Como instrumento para a realização da CPP-cor, esse trabalho utilizou-se de uma adaptação ao código referente a metodologia CPP-tri, implementado em linguagem R e disponibilizado por Silva (2016). O código referente a metodologia proposta nesta tese, encontra-se disponível na comunidade `R` (R Core Team, 2019).

O cálculo numérico das probabilidades de uma normal multivariada costuma ser um problema difícil. Para isto, o método utiliza uma transformação descrita em Genz (1992), que além de simplificar o problema, coloca em um formato que permite cálculos eficientes usando algoritmos numéricos de integração múltipla padrão. Esse procedimento está disponível na biblioteca `mvtnorm`, e permite que o cálculo das probabilidades normais multivariadas, detalhadas na seção 3.3.1, sejam moderadamente precisas e eficientemente computadas em problemas com até 1000 dimensões.

O objetivo dessa abordagem, é realizar a classificação supervisionada de i indivíduos em uma das r classes ordenadas e identificadas por q perfis, levando em consideração a relação existente entre as m variáveis observadas nos indivíduos. Para melhor descrever essa metodologia, considere um conjunto de dados observados em n indivíduos e relativos a m variáveis. Formalmente, representa-se esse conjunto da seguinte forma

Antes da metodologia ser aplicada, calcula-se em todas as classes ordenadas C_r, as média das observações obtida em cada variável j. Dessa maneira, o q-ésimo perfil representante (Y_{qm}) de uma classe genérica , é constituído por um vetor de médias. Formalmente, temos

$$Y_{qm} = [\bar{y}_{11}, \bar{y}_{12}, \ldots, \bar{y}_{qm}] \tag{11}$$

em que, $\bar{y}_{qm}$ é a média das observações na m-ésima variável do q-ésimo perfil.

Após definidos os perfis, considera-se para melhor identificar cada classe o relacionamento entre as variáveis que caracteriza cada individuo em cada classe. Para tanto, ancorou-se no cálculo das correlações entre as variáveis. Formalmente essas correlações são representadas pela matriz abaixo

$$\boldsymbol{\varrho} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \dots & 1 \end{pmatrix}. \tag{12}$$

em que , $\rho_{pp'}$ é o coeficiente de correlação entre as variáveis pp'.

Os perfis representativos das classes, são definidos de modo que as classes estejam efetivamente ordenadas. A restrição de monotonicidade, impõe que não existam classes altas com avaliações nos perfis menores que as obtidas por uma classe inferior. Para garantir esse comportamento entre as classes, foi utilizado um algoritmo de correção encontrado em Silva (2016).

3.3 Etapas do CPP-cor

As observações em $\boldsymbol{x_i^T}$, são assumidas como parâmetros de locação da distribuição normal multivariada, e vistos apenas como observações de m variáveis aleatórias que se comportam segundo essa função de distribuição conjunta. O mesmo é realizado com as médias $\bar{y}_{qm}$ que compõem os perfis de referência. Esse procedimento de "aleatorização"é denominado como transformação probabilística, proposto por Sant'Anna (2002).

Após as transformações nessas observações, cada indivíduo a ser classificado é representado por um vetor aleatório $\boldsymbol{X_i^T}$, e o mesmo acontece com as médias nos perfis de referência $\boldsymbol{Y_{qm}}$. Assim, tomamos cada indivíduo como uma variável aleatória m-dimensional e derivamos comparações probabilísticas entre os indivíduos e os perfis que representam as classes. O método CPP-cor envolve duas etapas.

3.3.1 O cálculo das probabilidades de desempenho

A probabilidade do indivíduo $\boldsymbol{X_i^T}$ possuir desempenho inferior e superior à classe C_r , identificada pelo perfil de referência $\boldsymbol{Y}_{qm}$ é dado por

$$\beta_{ir}^{-} = \int_{-\infty}^{\bar{y}_{1m}} \int_{-\infty}^{\bar{y}_{2m}} \dots \int_{-\infty}^{\bar{y}_{qm}} f_{\boldsymbol{X_i^T}}(\boldsymbol{t}) dt_1 dt_2, \dots dt_k \tag{13}$$

$$\beta_{ir}^{+} = \int_{\bar{y}_{1m}}^{\infty} \int_{-\bar{y}_{2m}}^{\infty} \dots \int_{\bar{y}_{qm}}^{\infty} f_{\boldsymbol{X_i^T}}(\boldsymbol{t}) dt_1 dt_2, \dots dt_k \tag{14}$$

sendo $f_{\boldsymbol{X}}(\boldsymbol{x})$, a função densidade de probabilidade da distribuição normal multivariada, apresentada em 2.4.

3.3.2 Alocação nas classes mais próximas

Por fim, os indivíduos serão alocados à classe em que a distância probabilística γ é mais próxima do valor nulo. Calcula-se as distâncias probabilísticas da seguinte maneira

$$\gamma = |\beta_{ir}^{+} - \beta_{ir}^{-}|$$

3.4 Metodologia do Estudo de Simulação

O problema de classificação monótona pressupõe sua utilização em dados nos quais existe a classificação de grupos, sendo que os referidos grupos devem apresentar entre si uma estrutura hierarquizada ordinal. Ou seja, os grupos devem possuir entre si ordenação por importância, natureza ou outro tipo de hierarquia de modo que seja possível a identificação ordinal de cada um deles.

Esta característica peculiar do problema de classificação monotônica torna uma tarefa já complexa, a obtenção de dados para estudos, ainda mais desafiadora, dado que disponibilidade de conjuntos de dados que apresentam esta característica fundamental, mesmo nos repositórios de dados de maior destaque, é bastante limitada. Deste modo, para avaliar a qualidade da metodologia proposta em relação ao desem-

penho da metodologia tradicional CPP-tri, faz-se necessária a utilização de estratégias de simulação de dados que permitam a comparação entre as abordagens metodológicas de interesse.

Ainda que opte-se pela simulação de dados, existem dois obstáculos a serem superados. O primeiro é a geração de grupos que garantidamente apresentem classificação ordinal. A segunda barreira é a garantia de existência de correlação entre as variáveis, uma vez que a metodologia proposta neste trabalho tem como hipótese que, na presença de correlação entre as covariáveis, a metodologia denominada CPP-cor apresenta resultados superiores à abordagem apresentada em Sant'Anna (2015). A seguir será apresentada a estratégia de simulação que garante a presença de ambas características nos dados simulados, que serão utilizados na aferição do desempenho tanto da metodologia clássica quanto da metodologia proposta.

As simulações e análises foram conduzidas em uma máquina equipada com processador Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz com 32gb de memória ram, utilizando o ambiente de programação e análise de dados R (R Core Team, 2019), em sua versão 3.6.1 para o sistema operacional Ubuntu 18.04.3 LTS.

Consideramos formalmente a matriz x, de dimensão $(n \times p)$, como uma representação dos conjuntos de dados simulados.

$$
\boldsymbol{X} = \begin{pmatrix}
x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\
x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{ip} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np}
\end{pmatrix}
$$

$$
= \begin{pmatrix}
\boldsymbol{x}_1^T \\
\boldsymbol{x}_2^T \\
\vdots \\
\boldsymbol{x}_i^T \\
\vdots \\
\boldsymbol{x}_n^T
\end{pmatrix} = \begin{pmatrix}
\boldsymbol{x}_{(1)} & \dots & \boldsymbol{x}_{(j)} & \dots & \boldsymbol{x}_{(p)}
\end{pmatrix}
$$

em que o vetor linha $\boldsymbol{x}_i^T$, representa o vetor p-dimensional de observações simuladas do i-ésimo indivíduo, e o vetor coluna $\boldsymbol{x}_{(j)}$ representa o vetor n-dimensional de observações simuladas correspondente a j-ésima variável, com $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, p$.

Conforme descrito na seção anterior, em função da baixa disponibilidade de dados que atendam às peculiaridades do problema, bem como para garantir um volume de dados suficientes para subsidiar uma comparação entre as metodologias CPP-tri e CPP-cor, foi realizado um estudo de simulação, com o intuito de gerar conjuntos de dados para aplicação das referidas metodologias e apuração de sua eficácia ao classificar os elementos em grupos. Considerando a estratégia de simulação descrita anteriormente, foram realizadas as seguintes oscilações nos parâmetros de simulação:

- Número de variáveis - d: 5, 8 e 12;
- Número de grupos ordenados - G: 5;
- distância entre grupos - δ: 0.25, 0.5, 0.75, 1;
- Intensidade de correlação (desconsiderando a diagonal principal, que é unitária):
  - Baixas - $100\% \in [-0.3, 0.3]$;
  - Moderadas - $100\% \in [-0.6, -0.3] \bigcup [0.3, 0.6]$;
  - Altas - $100\% \in [-1, -0.6] \bigcup [0.6, 1]$;
  - Custom 30-40-30 - $30\% \in [-0.3, 0.3]$; $40\% \in [-0.6, -0.3] \bigcup [0.3, 0.6]$ $30\% \in [-1, -0.6] \bigcup [0.6, 1]$;
  - Custom 50-30-20 - $50\% \in [-0.3, 0.3]$; $30\% \in [-0.6, -0.3] \bigcup [0.3, 0.6]$ $20\% \in [-1, -0.6] \bigcup [0.6, 1]$;
  - Custom 70-00-30 - $70\% \in [-0.3, 0.3]$; $30\% \in [-1, -0.6] \bigcup [0.6, 1]$;
  - Custom 60-30-10 - $60\% \in [-0.3, 0.3]$; $30\% \in [-0.6, -0.3] \bigcup [0.3, 0.6]$ $10\% \in [-1, -0.6] \bigcup [0.6, 1]$;

Tais variações foram inseridas de modo a garantir que uma diversidade interessante de cenários realistas fosse garantida, de modo a verificar as nuances de cada metodologia e os cenários nas quais cada uma delas se destaca. O número de variáveis foi incluído com o intuito de verificar se a dimensão do banco de dados influencia no desempenho dos métodos. A distância entre grupos visa checar a dificuldade de classificação de cada método quando os grupos encontram-se sobrepostos ou bem definidos.

Já as intensidades de correlação foram inseridas com o intuito de aferir o quanto a presença de correlação é capaz de influenciar o comportamento de cada método, sendo o parâmetro de maior destaque deste estudo de simulação, dado que a hipótese do estudo indica que o método de classificação apresenta um comportamento melhor se o cálculo das probabilidades considera a correlação entre as variáveis. O número de grupos foi fixado em 5 devido a um estudo piloto que indicou que este aspecto não influencia na qualidade de nenhuma das metodologias, portanto geraria esforço computacional inócuo.

Para uma comparação justa, a cada iteração do estudo de simulação, foi gerado um conjunto de dados considerando os parâmetros momentâneos do estudo e a partir deste conjunto de dados ambos os métodos foram aplicados a este conjunto, gerando

a classificação. Após a classificação realizada por cada método, foi aferida a acurácia de cada método, que será comparada entre os métodos, em função de cada um dos parâmetros.

Em função do tempo computacional elevado gasto na nova metodologia, devido à complexidade do cálculo da probabilidade conjunta, para cada combinação dos parâmetros δ, d e *intensidade de correlação*, foram realizadas 50 realizações, gerando um total de 8400 observações, sendo 4200 referentes a cada metodologia. Entende-se que o número de experimentos foi suficiente, uma vez que se apurará a acurácia em termos de média e mediana.

Os dados foram simulados por meio de distribuição normal multivariada, realizada por meio do pacote `mvtnorm`, biblioteca presente no software R (R Core Team, 2019) e proposto por (GENZ et al., 2020). Nesta biblioteca, a geração de dados depende basicamente de três parâmetros, a saber: tamanho da amostra, vetor de médias e matriz de correlações ou de covariâncias. Considere gerar um conjunto de dados com grupos ordenados de d variáveis e k grupos ordenadamente classificados.

Em relação à classificação monotônica, esta será obtida por meio da geração de vetores de médias com uma distância δ em relação a um vetor de médias inicial. Defina μ_0 o vetor de médias do grupo inicial tal que $\mu_0 \sim N(0, 1)$ sendo μ_0 um vetor de d entradas e Σ uma matriz de correlações de dimensão d. A metodologia utilizada para simulação da matriz de correlação $\mathbf{\Sigma}$, foi inspirada no método proposto por Marsaglia e Olkin (1984). Com base nestes elementos serão gerados dados por meio da distribuição normal multivariada, abreviada como $\boldsymbol{N}_p$.

A estratégia de geração de dados é descrita por meio do seguinte procedimento:

1. Gere aleatoriamente o vetor g com k valores obtidos aleatoriamente da distribuição uniforme discreta $U_D(50,\ 200)$[6] para definir o tamanho de cada grupo. ;
2. Gere μ_0 e Σ conforme descrito acima;
3. Gere o grupo de observações inicial G_1 com os seguintes parâmetros: $g[1]$ observações da distribuição $\boldsymbol{N}_p(\mu_0,\ \Sigma)$;
4. Calcule o vetor σ contendo os desvios-padrão de cada coluna de variáveis de G_1;

[6] A utilização da distribuição uniforme supracitada visa variação do tamanho dos grupos aleatoriamente em cada amostra gerada.

5. Defina uma distância entre grupos δ;
6. Gere o grupo de observações G_2 com os seguintes parâmetros: $g[2]$ observações da distribuição $\boldsymbol{N}_p(\mu_0 + \delta\sigma,\ \Sigma)$;
7. Gere o grupo de observações G_3 com os seguintes parâmetros: $g[3]$ observações da distribuição $\boldsymbol{N}_p(\mu_0 + 2\delta\sigma,\ \Sigma)$;
8. Proceda sucessivamente até a geração do grupo de observações G_k com os seguintes parâmetros: $g[k]$ observações da distribuição $\boldsymbol{N}_p(\mu_0 + (k-1)\delta\sigma,\ \Sigma)$;
9. Agregue os grupos $G_1, G_2, \cdots, G_k$ em um único conjunto de dados.

A estratégia descrita acima garante que os grupos $G_1, G_2, \cdots, G_k$ possuem uma hierarquia ordenada em função das distâncias de δ nos vetores de média, além da possibilidade de controlar a intensidade da distância entre cada grupo por meio do parâmetro δ. Esta possibilidade será importante para aferir se os métodos em análise são sensíveis à sobreposição de grupos. Além disto garante a incorporação da correlação entre as variáveis, principal característica em estudo neste trabalho.

Definida a estratégia de geração dos dados, o estudo de simulação se dará por meio de oscilações nos parâmetros d, k, δ, além da intensidade das correlações componentes da matriz $\boldsymbol{\Sigma}$.

3.5 Métricas de avaliação do modelo

A acurácia pode ser definida como a proporção de acertos do modelo dentro do total de tentativas que o modelo executou. Considere a seguinte matriz de confusão de um modelo multiclasses:

	Observado				
Predito	Classes	C_1	C_2	$\cdots$	C_N
	C_1	c_{11}	c_{12}	$\cdots$	c_{1n}
	C_2	c_{21}	c_{22}	$\cdots$	c_{2n}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	C_N	c_{n1}	c_{n2}	$\cdots$	c_{nn}

na qual C_i representa cada uma das classes observadas nos dados e c_{ij} representa o número de elementos da classe i classificados pelo modelo na classe j. Considera-se uma classificação correta os valores do tipo c_ii, ou seja, aqueles cujas

classes observada e predita coincidem. Assim, pode-se calcular a acurácia por meio da seguinte fórmula:

$$Acc = \frac{\sum_{i=1}^{n} c_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \tag{15}$$

Um modelo acurado apresenta valores significativos na diagonal principal da matriz de confusão e valores pequenos nas porções triangulares superior e inferior da mesma. Por se tratar de uma proporção, a acurária pertence ao intervalo $[0, 1]$, sendo desejável valores próximos de um para tal medida. Deste modo, a acurácia será a medida utilizada para a comparação do desempenho do método proposto em relação à metodologia CPP-tri.

3.6 Teste de Wilcoxon

O objeto de comparação entre as metodologias será a acurácia obtida por cada metodologia aplicada em um mesmo conjunto de dados, o que compõe um experimento pareado. Entretanto, devido à natureza não normal da acurácia, cujos valores pertencem ao intervalo $[0, 1]$, não será possível a utilização de metodologias paramétricas, tais como o teste t pareado. Por este motivo será utilizado o teste de Wilcoxon para comparação entre as medianas.

O teste de Wilcoxon (WILCOXON, 1945), é um teste de hipóteses não paramétrico utilizado para comparação de duas amostras relacionadas, pareadas ou de medidas repetidas para verificar se o ranqueamento das amostras difere, podendo assim ser utilizado como uma alternativa ao teste t pareado de Student. Este teste verifica se duas amostras advém da mesma população.

Sejam duas amostras pareadas $x_1, x_2, \cdots, x_n$ e $y_1, y_2, \cdots, y_n$. Caso as amostras sejam oriundas de uma mesma população, espera-se que a mediana da diferença $d_= x_i - y_i$ seja igual a zero, ou seja, desejamos testar $H_0 : d_i = 0$ vs $H_1 : d_i \neq 0$. Para a realização do teste, devem ser seguidos os seguintes procedimentos:

1. Calcular cada diferença pareada $d_i = x_i - y_i$;
2. Ranquear as diferenças d_i independente de seu sinal;

3. Atribuir ao ranking atribuído o sinal de d_i;
4. Calcular as estatísticas W^+ - soma dos rankings positivos e W^- - soma dos rankings negativos;
5. Escolher $W = \min(W^-, W^+)$;
6. Utilizar a tabela dos valores críticos do teste para concluir sobre o teste realizado ou utilizar a aproximação pela distribuição normal com média $\mu_W = \frac{n(n+1)}{4}$ e $\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

Detalhes adicionais, como a solução de empates nos pares podem ser obtidos em Sprent e Smeeton (2000).

4 RESULTADOS E DISCUSSÃO

Neste capítulo, são reportados e discutidos os resultados obtidos a partir da aplicação dos métodos de classificação CPP-Tri e CPP-cor em bases de dados simuladas e obtidas a partir do repositório UCI(DUA; GRAFF, 2017) e Github (GITHUB, 2019). Inicialmente serão apresentados alguns resultados referentes ao tempo de execução dos experimentos e na sequência sera apurada a qualidade de cada metodologia em função de sua acurácia em classificar os indivíduos corretamente.

4.1 Análise do tempo computacional

Em relação ao tempo de execução, conforme esperado, a metodologia CPP-tri apresentou um tempo de execução menor se comparado com a metodologia CPP-cor. Enquanto o CPP-tri executou os procedimentos com tempo médio de 5,02 segundos, o CPP-cor apresentou um tempo médio de execução de 123.56 segundos. Tal fato já era esperado, dada a complexidade computacional do cálculo de probabilidades conjuntas.

Tabela 4: Tempo de execução em minutos de cada método em função do número de variáveis.

	Tempo de Execução			
	CPP-TRI		CPP-cor	
var	mean (sd)	min - max	mean (sd)	min - max
5	4.429 (0.834)	2.321 - 8.426	11.800 (2.501)	6.21 - 22.241
8	4.956 (0.940)	2.580 - 8.655	59.786 (12.183)	28.216 - 99.65
12	5.704 (1.195)	2.716 - 14.888	299.107 (57.982)	146.521 - 522.347

A Tabela 4 apresenta o impacto do crescimento do banco de dados no tempo de execução de cada metodologia em função do número de variáveis. Observa-se que o tempo de execução oscila bastante em função do número de variáveis que compõe o conjunto de dados, devido à anteriormente mencionada complexidade dos cálculos de probabilidades conjuntas.

Em relação a distÂncia entre os grupos, a Tabela 5 indica que a conFiguração dos grupos em termos da distância entre si não impacta diretamente no tempo médio

Tabela 5: Tempo de execução em minutos de cada método em função da distância entre grupos.

	Tempo de Execução			
	CPP-TRI		CPP-cor	
δ	mean (sd)	min - max	mean (sd)	min - max
0.25	5.059 (1.096)	2.53 - 9.58	131.381 (140.669)	6.407 - 522.347
0.5	4.979 (1.056)	2.46 - 8.735	124.098 (130.493)	6.21 - 505.646
0.75	4.986 (1.155)	2.321 - 14.888	119.877 (124.977)	6.757 - 456.193
1	5.094 (1.203)	2.537 - 14.847	118.901 (124.029)	6.423 - 438.236

de execução. Ressalta-se que apesar de apresentar uma média de tempo de execução maior, o tempo de execução da metodologia CPP-cor não é proibitivo, sobretudo se o método conseguir garantir resultados superiores à metodologia padrão.

Havendo consistência nos resultados apresentados, gerando ganhos na qualidade de classificação, o tempo de execução torna-se irrelevante em detrimento a uma maior acurácia na classificação final dos indivíduos.

4.2 Análise da acurácia de cada modelo

A verificação da acurácia da classificação de cada um dos métodos se dará por meio da acurácia. A acurácia é a proporção de classificações corretas dentre as tentativas de classificação do método em estudo. Por se tratar de uma medida que não apresenta distribuição normal e de aplicação de duas técnicas distintas em um mesmo conjunto de dados, será utilizado o teste pareado de Wilcoxon (WILCOXON, 1945), com o objetivo de verificar se a acurácia mediana entre as metodologias CPP-TRI e CPP-cor são iguais.

A primeira análise trata do desempenho geral dos dois métodos no estudo de simulação. A análise gráfica apresentada na Figura 6 demonstra indícios de uma superioridade da metodologia CPP-cor. Estes indícios são comprovados por meio do teste pareado de Wilcoxon, apresentado na Tabela 6, ao nível de significância de 1%. Deste modo, conclui-se que, considerando o cenário geral, a metodologia CPP-cor apresenta acurácia de classificação superior se comparada à metodologia CPP-TRI

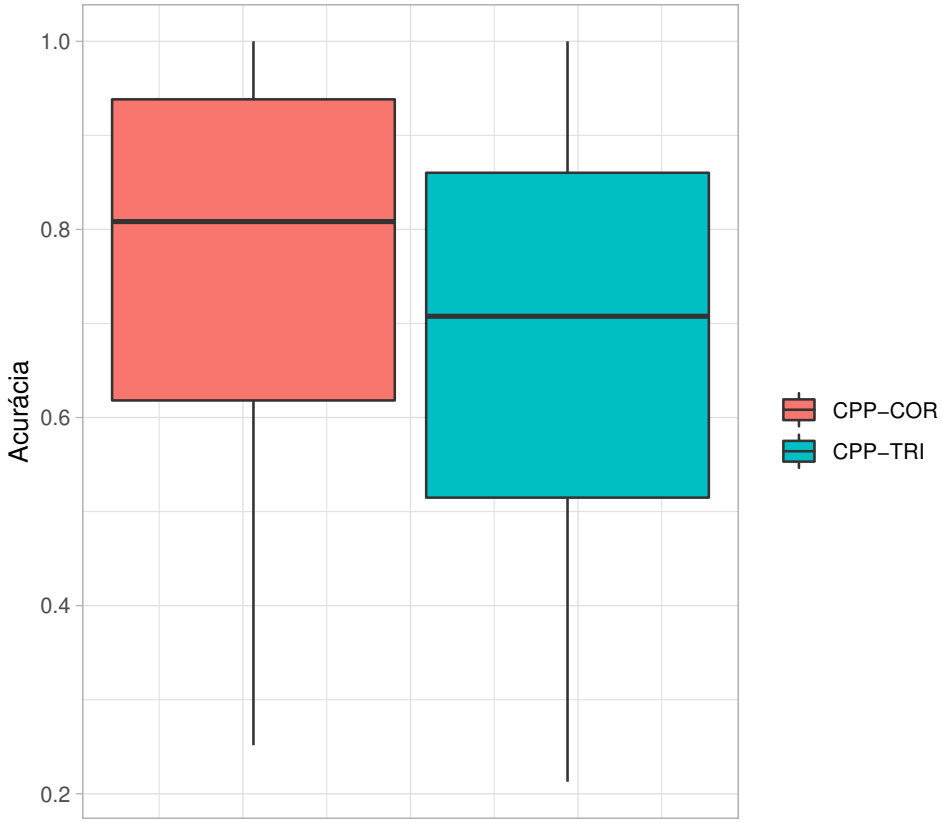

Figura 6: Boxplot - Acurácia geral dos métodos CPP-TRI e CPP-cor no estudo de simulação

Tabela 6: Acurácia de cada método no caso geral.

Acurácia		
Mediana		
CPP-TRI	CPP-cor	valor-p
0,076	0,8083	0

Após a verificação do desempenho geral dos métodos, foi realizada uma análise detalhada da acurácia em relação às características consideradas influentes no processo de classificação dos indivíduos nos grupos, com o intuito de verificar se existem padrões que definem casos nos quais as metodologias são superiores entre si.

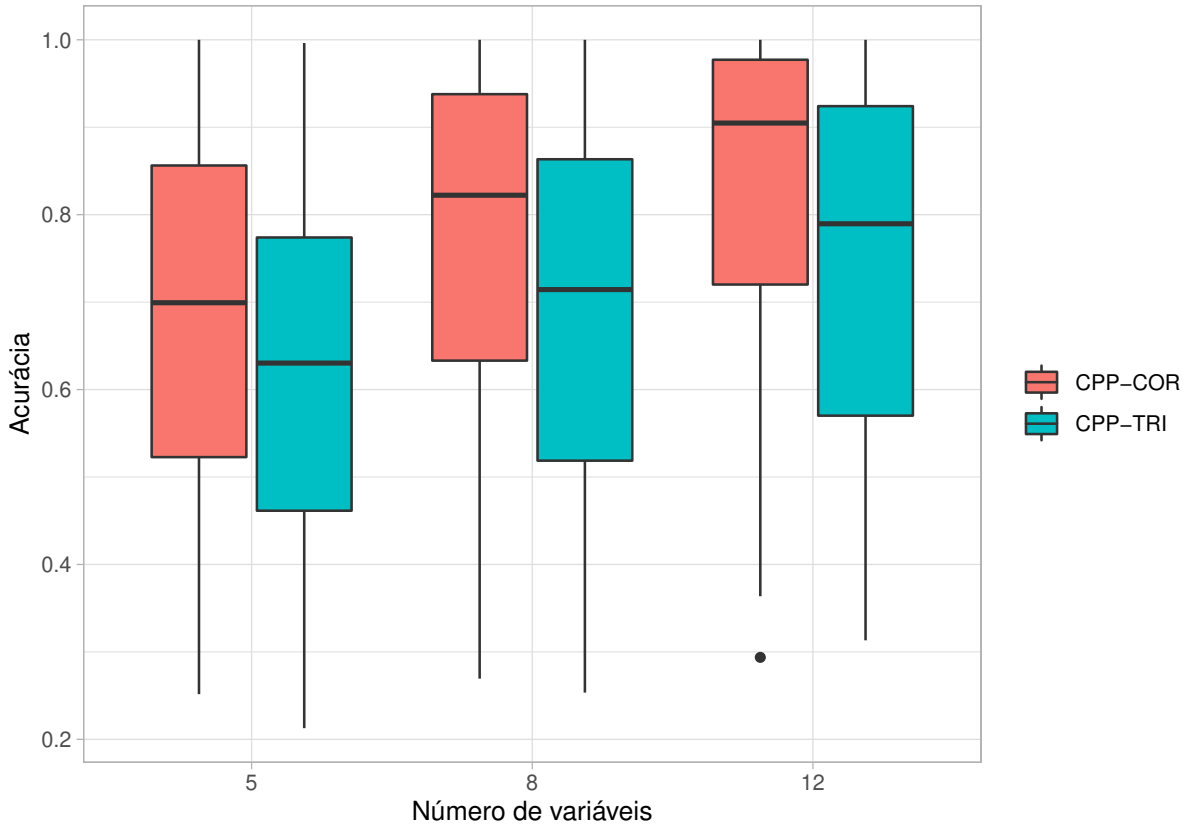

Figura 7: Boxplot - Acurácia de cada método em função do número de variáveis.

Primeiramente, foi verificada a acurácia de cada método em relação ao número de variáveis. Assim como no caso geral, observa-se que o CPP-cor apresenta desempenho superior em todos os cenários se levado em consideração o número de variáveis do conjunto de dados. A análise gráfica (Figura 7) deixa explícita a vantagem do método CPP-cor em relação à abordagem tradicional, o que é corroborado pelo teste de hipóteses apresentado na Tabela 7, que aponta diferença significativa entre todos os grupos ao nível de 1% de significância.

Note que com o aumento no número de variáveis, existe um acréscimo na diferença entre as medianas. Este acréscimo indica que o método CPP-cor apresenta um ganho de qualidade em relação ao método CPP-cor conforme cresce a dimensão do conjunto de dados.

Um outro aspecto importante no estudo de simulação é checagem em relação à

Tabela 7: Acurácia de cada método em função do número de variáveis.

Acurácia			
	Mediana		
Nº de Variáveis	CPP-TRI	CPP-cor	valor-p
5	0,6303	0,6993	0
8	0,7143	0.8222	0
12	0,7896	0.9048	0

distância entre os grupos. O objetivo é verificar se alguma das metodologias apresenta variações em relação à qualidade em função da sobreposição dos grupos, ou seja, investigar se os métodos apresentam queda de qualidade quando os grupos não são completamente heterogêneos.

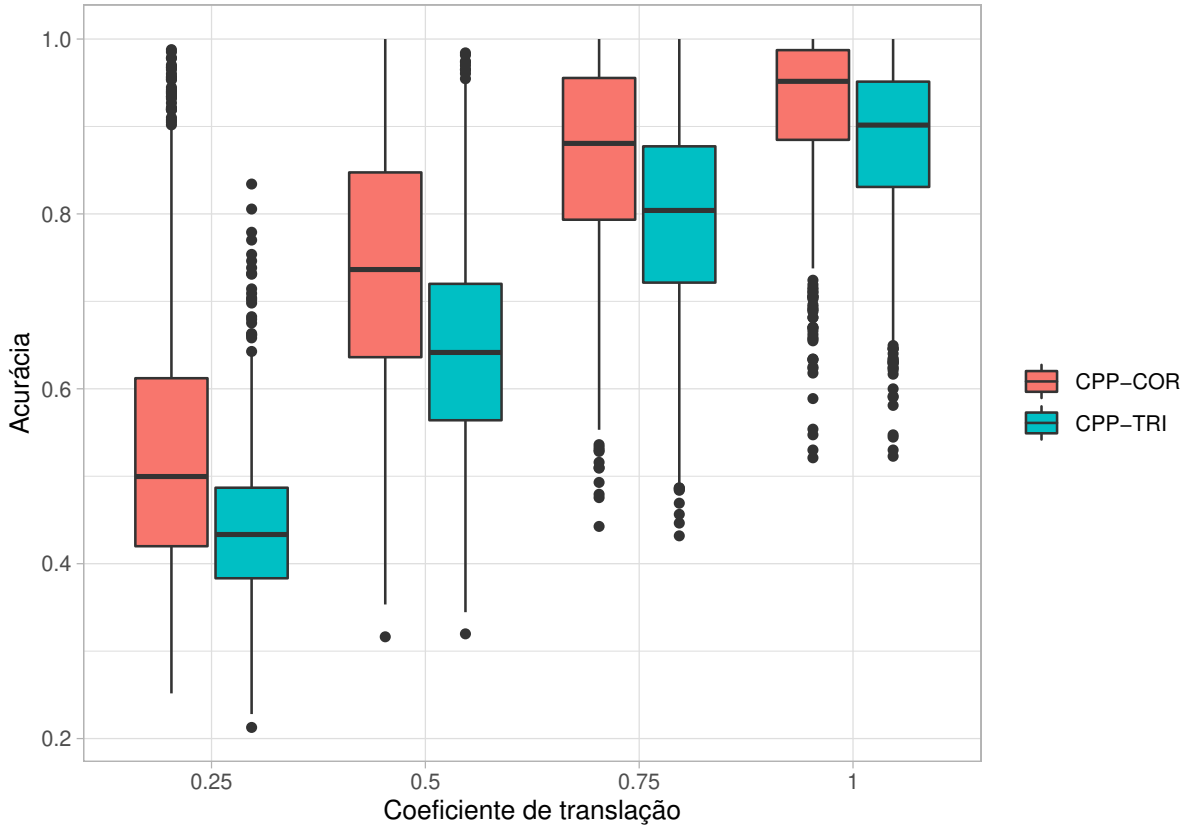

Figura 8: Boxplot - Acurácia de cada método em função da distância entre grupos.

A análise descritiva dos resultados (Figura 8) apresenta indícios de que quanto mais heterogêneos os grupos, maior a acurácia dos métodos de classificação, o que é esperado. Observa-se também que o CPP-cor mais uma vez apresentou resultados

superiores em relação ao CPP-cor, conforme conclui-se ao nível de 5% de significância por meio dos testes de hipóteses apresentados na Tabela 8.

Tabela 8: Acurácia de cada método em função da distÂncia entre grupos.

Acurácia			
	Mediana		
Distância entre grupos (δ)	CPP-TRI	CPP-cor	valor-p
0.25	0.4334	0.4997	0
0.50	0.6416	0.7364	0
0.75	0.8040	0.8806	0
1.00	0.9015	0.9516	0

O estudo da acurácia dos métodos em relação à intensidade da correlação será dividido em duas partes. A primeira terá como base conjuntos de dados nos quais a intensidade de correlação foi definida como *Baixa*, *Moderada* e *Alta*. A segunda se dará por meio de dados gerados via matrizes de correlações customizadas.

O estudo dos métodos com base nas correlações baixas, moderadas e altas visa identificar o comportamento das metodologias em casos específicos, que dificilmente serão observados na prática, porém agregam valor teórico devido à possibilidade de identificar o quanto a correlação impacta em cada método. Observa-se na Figura 9 que na ausência de correlação, ou seja, no caso de dados gerados de matrizes de correlação de baixa intensidade de correlação, os métodos apresentam comportamento bastante semelhante. Este fato já era esperado, uma vez que da teoria de probabilidade sabe-se que a probabilidade conjunta de variáveis aleatórias independentes se fatora no produto de suas marginais. Assim, haveria diferenças mínimas no cálculo das probabilidades via distribuições marginais ou conjuntas.

Porém, quando se acrescenta uma matriz de correlações moderadas no processo de geração dos dados, já se percebe uma diferenciação na acurácia mediana de cada metodologia, sendo que o CPP-cor se destaca. A presença de correlação faz com que o cálculo da probabilidade conjunta agregue mais informação ao método, fazendo com que este descrimine melhor os grupos. A diferença entre a qualidade dos métodos fica ainda mais evidente quando se gera dados com variáveis altamente correlacionadas, conforme pôde-se observar. Os testes de hipóteses apresentados nas

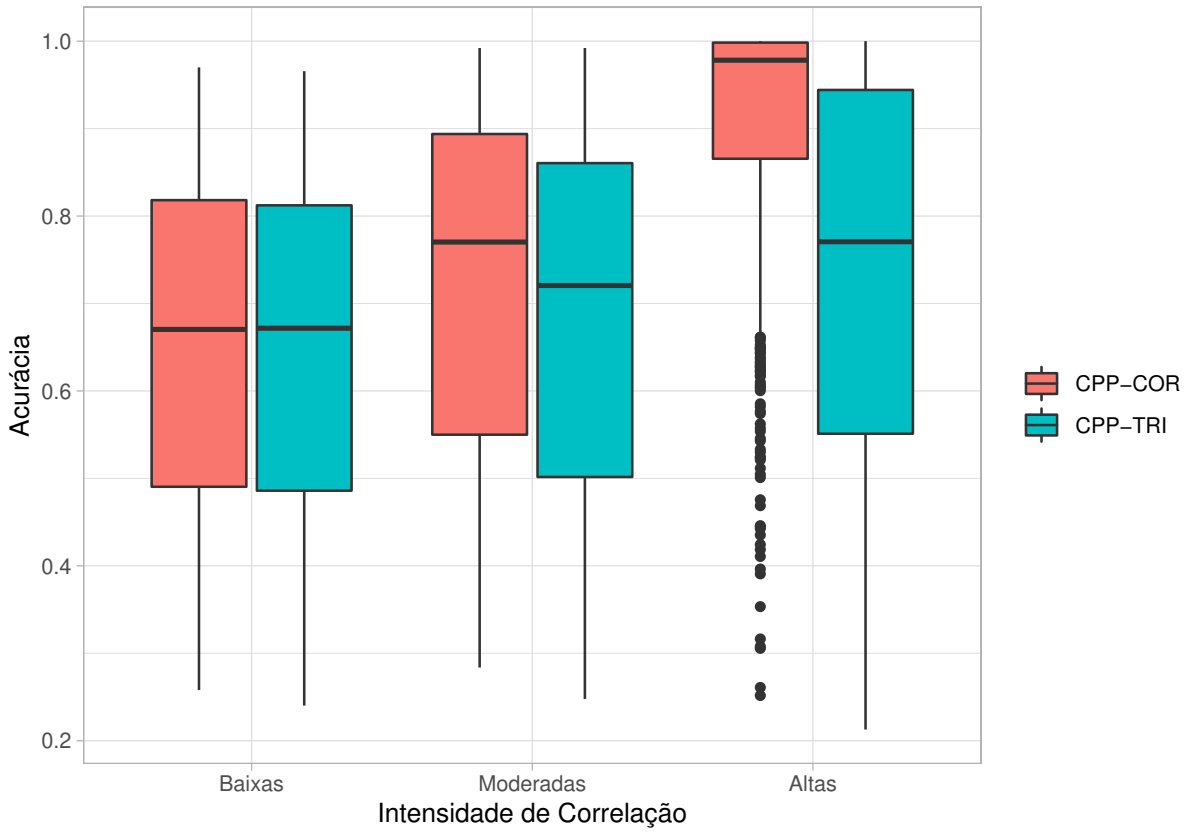

Figura 9: Boxplot: Acurácia de cada método em função da intensidade de correlação - Intensidades constantes

três primeiras linhas da Tabela 9 confirmam as impressões obtidas na análise visual.

A segunda parte do estudo de intensidade de correlação visa simular a aplicação dos métodos em cenários mais aproximados da realidade, na qual se espera que os dados possuam em suas matrizes de correlação uma mescla entre correlações baixas, médias e altas. Foram simulados quatro cenários, previamente descritos, em que se busca variações que podem ser encontradas em conjuntos de dados reais.

A análise descritiva dos resultados (Figura 10) mais uma vez aponta para a superioridade do método CPP-cor quando da presença de correlações significativas entre as variáveis. É interessante notar que a distância entre a qualidade dos métodos apresenta decréscimo do primeiro ao terceiro gráfico, crescendo novamente no quarto gráfico. Este comportamento ocorre devido à proporção de correlações altas, com módulo superior à 0.6, parte de 30% para 20% e depois para 10% nos três primeiros cenários, aumentando para 30% no último deles. Este comportamento aponta que o CPP-cor, que apresentou uma eficácia superior em relação ao CPP-TRI, apresenta desempenho ainda melhor quando há alta correlação entre os dados, informação des-

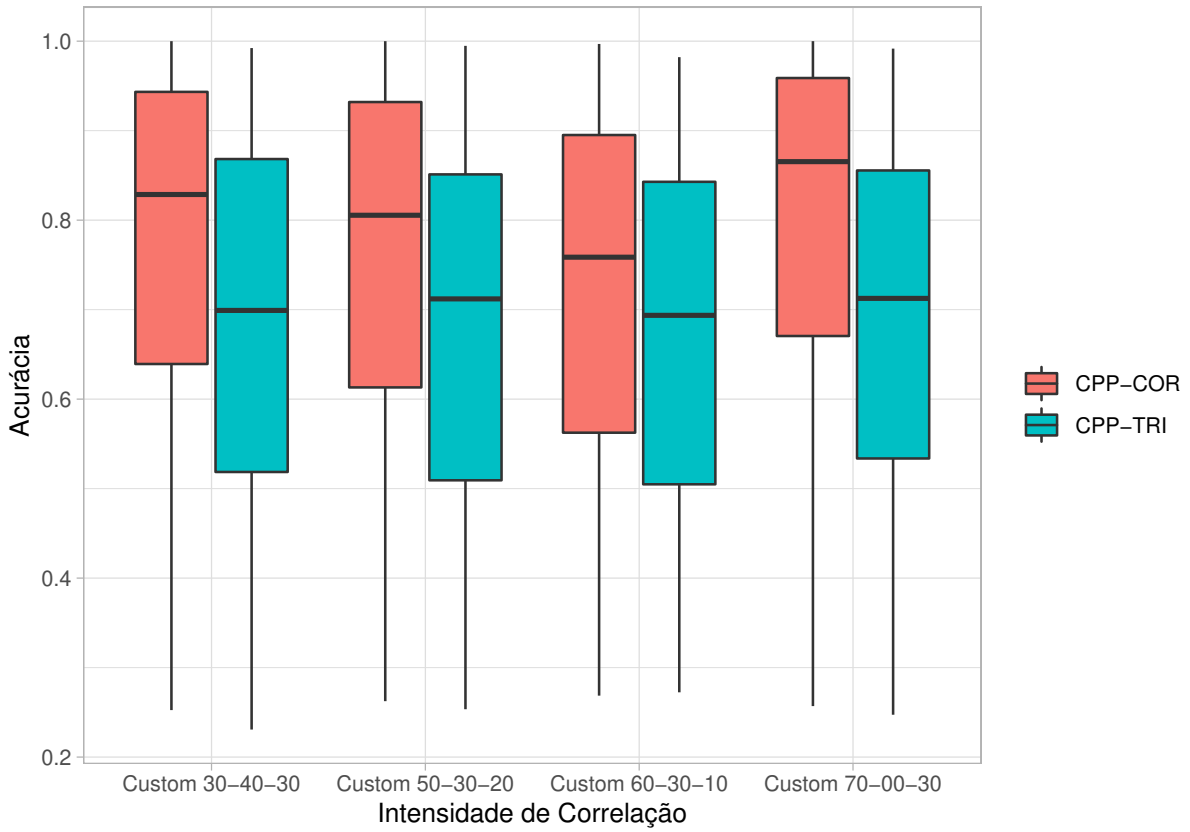

Figura 10: Boxplot: Acurácia de cada método em função da intensidade de correlação - Intensidades customizadas

presada pela metodologia tradicional.

Os testes de hipóteses apresentados nas 4 últimas linhas da Tabela 9 corroboram os resultados visuais apresentados, indicando que de fato há superioridade de qualidade do CPP-cor nos cenários simulados.

4.3 Aplicação das técnicas em Dados Reais

Os resultados obtidos por meio do estudo de simulação apresentam evidencias de que a metodologia proposta neste trabalho gera resultados bastante satisfatórios quando o conjunto de dados contém correlação entre as variáveis que o compõe. Entretanto, a aplicação em conjuntos simulados por muitas vezes não consegue explorar todas as nuances presentes em um conjunto de dados reais, pois independente da complexidade dos cenários construídos para a simulação, esta se limita a uma representação simplificada da realidade.

Para submeter a metodologia proposta ao crivo da realidade, novamente as técnicas tradicional e proposta serão comparadas, desta vez por meio da aplicação em

Tabela 9: Acurácia de cada método em função da intensidade de correlação.

Acurácia			
	Mediana		
Intensidade de Correlação	CPP-tri	CPP-cor	valor-p
Baixa	0,6717	0,6703	0,8948
Moderada	0,7206	0,7703	0,0002
Alta	0,7706	0,9783	0
Custom 30-40-30	0,6991	0,8286	0
Custom 50-30-20	0,7120	0,8055	0
Custom 60-30-10	0,6936	0,7586	0
Custom 70-00-30	0,7126	0,8654	0

um conjunto de dados já consagrado, sobretudo em estudos de regressão, o conjunto *Boston Housing Price*.

Como se trata de problema de classificação ordenada, inicialmente buscou-se uma estratégia de classificação dos dados em conjuntos de forma a apresentar certa hierarquia. Conforme descrito no capítulo 3, os dados provém de um estudo econométrico sobre o mercado imobiliário, sendo portanto a variável **MEDV**, valor médio das casas ocupadas pelos proprietários, uma importante variável considerada aqui nesse trabalho como a classe. Neste sentido, considerando uma possível correlação entre as variáveis, é lógica a pressuposição de que as demais variáveis estão correlacionadas ao valor médio das residências. Portanto, esta foi a variável utilizada para a classificação dos grupos, que foram delimitados por meio dos quartis da referida variável. Casas com preço médio pertencentes ao primeiro quartil foram agrupadas na classe 1, enquanto as com preço médio no segundo quarti foram agrupadas no quartil 2 e assim sucessivamente até se obter a classificação dos domicílios em quatro grupos. Deste modo, espera-se que os grupos apresentem uma classificação ordinal.

Após a classificação dos grupos, considerando que a metodologia tem como pressuposto a distribuição normal das variáveis individualmente, foram retiradas variáveis *dummy* e também variáveis pertenciam ao intervalo $[0,\ 1]$, como por exemplo as variáveis **B** e **ZN**. Após a remoção das referidas informações, foram mantidas as variáveis **NOX**, **CRIM**, **PTRATIO**, **MEDV**, **RM** e **DIS**. A seguir algumas estatísticas des-

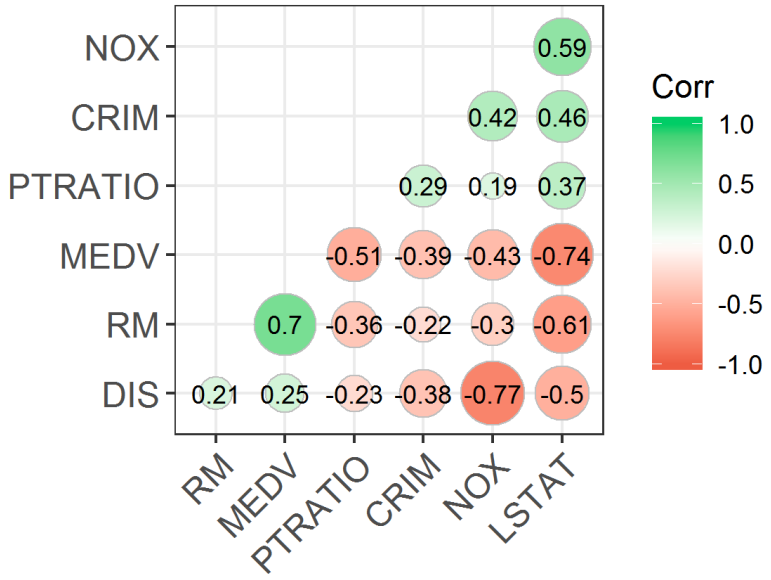

Figura 11: Correlograma do banco de dados Boston Housing

critivas do conjunto de dados resultante são apresentadas na Tabela 10.

Tabela 10: Estatísticas descritivas do conjunto de dados Boston Housing Price

	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
Variável	Média	D. Padrão	Média	D. Padrão	Média	D. Padrão	Média	D. Padrão
CRIM	7,75	14,21	0,21	3,35	0,13	2,24	0,09	2,19
NOX	0,68	0,09	0,54	0,10	0,49	0,10	0,47	0,08
RM	5,95	0,59	5,96	0,32	6,29	0,45	6,98	0,71
DIS	1,87	1,17	3,38	2,14	3,95	2,05	3,98	2,12
PTRATIO	20,20	1,77	19,60	1,71	18,50	1,83	17,40	2,15
LSTAT	19,92	5,97	13,09	4,42	9,09	3,93	5,23	3,08
MEDV	281,40	62,47	407,40	24,25	485,10	22,79	691,95	165,39

Nota-se que em todas as variáveis existe ordenação entre as médias dos grupos, seja crescente ou decrescente, no sentido do grupo um ao grupo quatro, indicando que a estratégia de agrupamento adotada logrou êxito na classificação ordenada dos grupos. Uma segunda preocupação é a presença de correlação nos dados em estudo, devido à hipótese do estudo de que o método proposto atua de forma superior à metodologia tradicional na presença de correlação. Para tanto, foi gerado o correlograma dos dados em estudo, apresentado na Figura 11.

Ao se analisar o correlograma, nota-se a presença de valores de correlação em faixas baixas, moderadas e altas, tanto positivas quanto negativas. Adicionada esta característica às supracitadas, conclui-se tratar de um conjunto de dados ideal para a

apuração da qualidade das metodologias em estudo.

De posse de um conjunto adequado à aplicação dos dados, procedeu-se à aplicação das metodologias para investigação do desempenho de cada uma delas. Primeiramente serão apresentados os resultados da metodologia CPP-tri. Observa-se que, de acordo com os resultados apresentados na Tabela 11, a metodologia CPP-tri, aplicada ao conjunto de dados *Boston Housing Prices* apresentou uma acurácia de 0,6324, ou seja conseguiu alocar corretamente 63,24% dos elementos em suas devidas classes.

Tabela 11: Matriz de Confusão - CPP-tri

	Classe Observada			
Classificação	1	2	3	4
1	98	19	2	0
2	28	73	28	3
3	1	36	53	25
4	0	1	43	96
Acurácia	0,6324			

Em contrapartida, de acordo com o disposto na Tabela 12, a metodologia proposta apresentou uma acurácia superior, no valor de 0,7272, ou seja, classificou devidamente 72,72% dos indivíduos, obtendo um desempenho aproximadamente 15% superior à metodologia tradicional de classificação, o que representa uma melhoria significativa dos resultados obtidos.

Tabela 12: Matriz de Confusão - CPP-cor

	Classe Observada			
Classificação	1	2	3	4
	110	21	0	0
	17	75	21	0
	0	33	80	21
	0	0	25	103
Acurácia	0,7272			

4.4 Pacote CPP-cor

Foi desenvolvido um pacote em linguagem R para a execução das metodologias testadas neste trabalho, com o objetivo de facilitar a aplicação das técnicas de classificação ordenada por parte dos usuários do método.

A motivação responsável pelo desenvolvimento dessa biblioteca, está totalmente relacionada a possibilidade de fornecer um instrumento com implementação em código aberto, que proporcione a replicabilidade da metodologia investigada nesta tese. Visto que, há uma extrema necessidade de áreas do conhecimento atuarem de forma transversal na solução de diversos problemas multidisciplinares, em especial, o da classificação monotônica de dados correlacionados. Todos os algoritmos em linguagem R referentes a metodologia e pacote proposto encontram-se disponível na biblioteca `cppcor` (RIBEIRO et al., 2020) no repositório CRAN (R Development Core Team, 2019). Os algoritmos desenvolvidos basearam-se no código em R que encontram-se disponível em Silva (2016).

O nome da biblioteca faz alusão à metodologia proposta por (SANT'ANNA, 2015), denominada como Probabilistic Composition of Preference. A sigla CPP-cor, referencia a metodologia definida nesse trabalho, intitulada de composição probabilística correlacionadas.

Até a publicação dessa tese, a biblioteca disponibiliza uma função nomeada como *CPP-cor*(12). Dois argumentos estão disponíveis para essa função, o primeiro consiste em receber todo o conjunto de dados, ou como descrito **dataset** (Figura 13). Esse, deve reservar a última coluna para possíveis valores da classe . Caso isso não aconteça, uma mensagem de erro será informada.

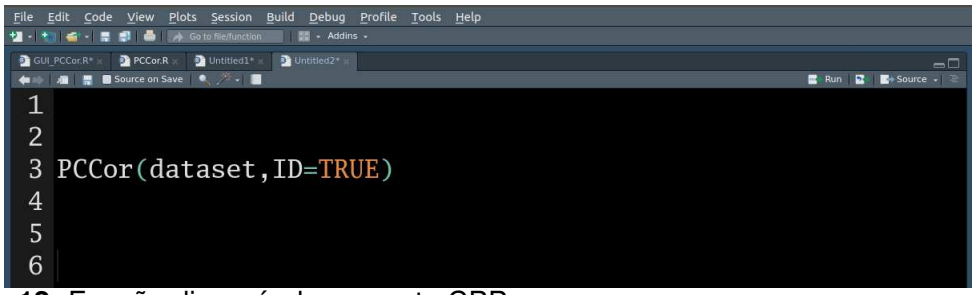

Figura 12: Função disponível no pacote CPP-cor.

O argumento **ID** faz referência a suposição de independência entre as variáveis, que consiste em uma condição necessária para composição probabilística desenvol-

vida pelo estatístico Sant'Anna (2015). Esse argumento pode receber dois valores lógicos, **TRUE** e **FALSE**. O valor **TRUE**, permite que a composição das informações contidas nas variáveis seja realizada de forma independente , em termos probabilísticos, essa composição se dá por meio do produto dessas informações (referencia da fórmula da composição). Vale salientar que nessa condição tem-se a aplicação do método CPP-tri (SANT'ANNA; COSTA; PEREIRA, 2015). Por fim , o valor **FALSE** possibilita compor as informações levando em consideração o relacionamento entre as variáveis que descrevem os indivíduos.

A biblioteca fornece um conjunto de dados para exemplo. A Figura 13 apresenta a aplicação da função **CPP-cor** a esse conjunto.

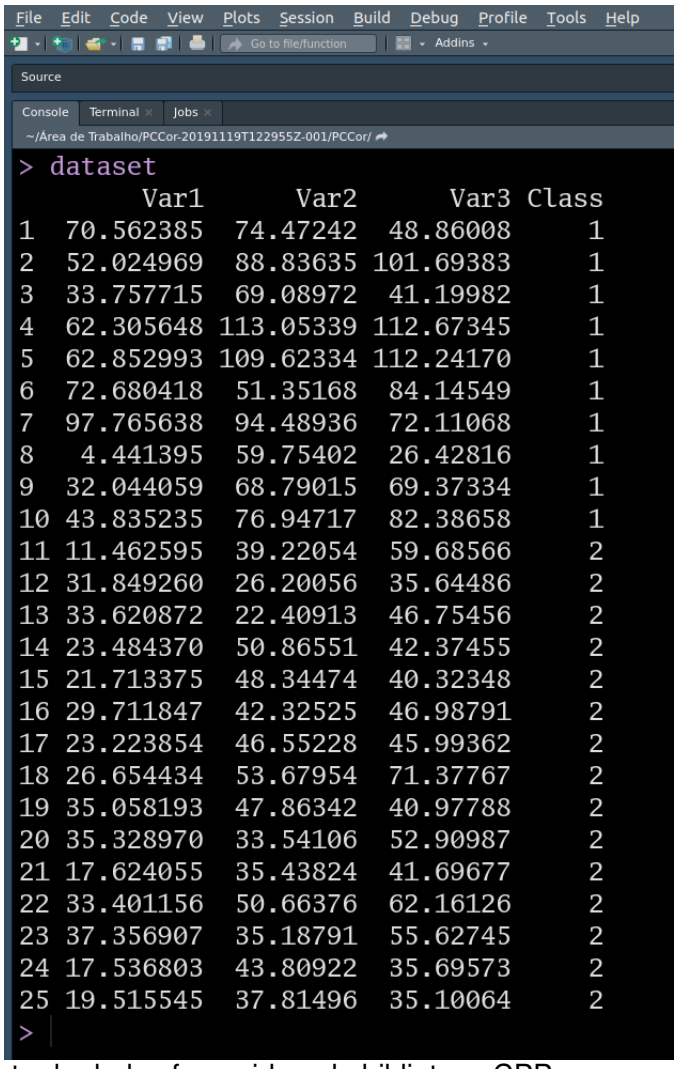

Figura 13: Conjunto de dados fornecido pela biblioteca CPP-cor.

Como o foco para essa função é suportar problemas de classificação ordinal de dados correlacionados, a compilação dessa função apresenta um saída (*output*)

estruturada com os resultados das medidas usuais para medir o despenho de métodos de classificação (14).

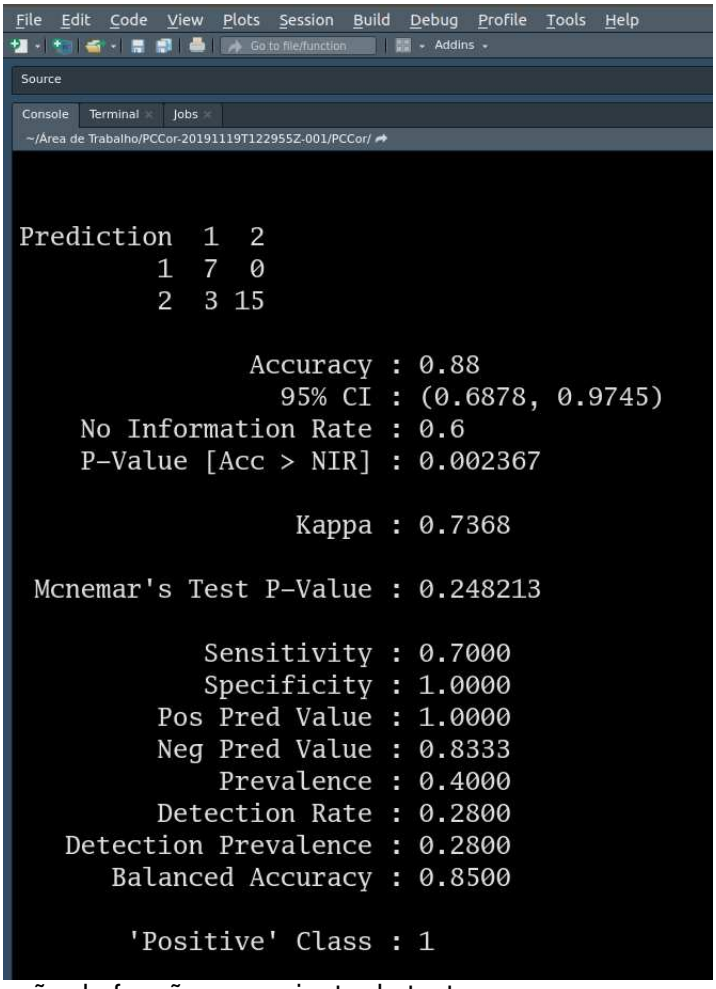

Figura 14: Aplicação da função no conjunto de teste.

4.5 Sumário

Neste capítulo foram apresentados resultados de estudos de simulação e da aplicação das metodologias CPP-tri e CPP-cor em conjuntos de dados reais. Após criteriosa análise dos resultados obtidos, observou-se que na ausência de correlações significativas, ambas as metodologias apresentam desempenho bastante semelhante. Entretanto, quando é incluída a correlação no processo de geração das variáveis ou é constatada a presença de correlação em conjuntos de dados reais, a metodologia proposta apresenta desempenho superior se comparada ao método CPP-tri.

Interessante notar também algumas das conclusões obtidas no estudo de simulação, como por exemplo a melhora do desempenho do método CPP-cor quando

ocorre o acréscimo no número de variáveis e também no percentual de correlações altas entre as variáveis. Por outro lado, a sobreposição menor ou maior entre os grupos aparentemente não gera um aumento na distância existente entre o desempenho dos dois métodos em estudo, indicando que ambos os métodos são capazes de classificar de forma ordinal os grupos de dados, independente de possíveis sobreposições das classes.

5 CONCLUSÕES

Conforme pôde se observar ao longo deste trabalho, a capacidade de ordenação é uma das características que auxiliou no sucesso da humanidade enquanto sociedade e sua presença é cada vez mais marcante, seja em aplicações cotidianas bem como em utilizações mais sofisticadas, como sistemas de recomendação, análise de riscos e inteligência artificial. Num mundo onde o volume de dados gerados pelas diversas plataformas de informação, o desenvolvimento de metodologias capazes de modelar a interação entre as componentes destes grandes conjuntos de dados é fundamental.

Neste sentido, a metodologia apresentada nesta tese é fundamentada na metodologia CPP-tri, proposta por Sant'Anna, Costa e Pereira (2015). A CPP-cor apresenta resultados superiores que a metodologia tradicional, uma vez que ela considera justamente a interação entre as variáveis presentes no conjunto de dados por meio da correlação, para composição das classes ordenadas.

A metodologia CPP-cor se mostrou tão eficiente quanto a metodologia CPP-tri em conjuntos de dados nos quais não há correlações significativas entre as variáveis que compõe as bases simuladas. Entretanto, quando as correlações foram incorporadas à base de dados, percebeu-se uma sensível diferenciação entre a acurácia dos métodos, com vantagem para a metodologia proposta neste trabalho. Por fim, o resultado obtido por meio da aplicação dos métodos no conjunto de dados reais *Boston Housing Price* corroboraram com aqueles obtidos pelo estudo de simulação apresentado nesta tese, fortalecendo a convicção de que a inclusão da correlação das variáveis na classificação dos elementos em classes ordenadas conferiu vantagem à nova metodologia em detrimento à metodologia CPP-tri.

Deste modo, pode-se concluir que houve êxito na inclusão da correlação como componente do cálculo das probabilidades de preferência utilizadas na classificação ordenada dos dados em estudo. Além deste êxito, destaca-se o desenvolvimento do pacote para o software R para a utilização tanto do método proposto, nomeado CPP-cor quanto da metodologia CPP-tri, o que permitirá a difusão da técnica e fomentará o desenvolvimento de novos estudos na área de classificação ordenada.

Como sugestão de trabalhos futuros, pretende-se desenvolver uma aplicação baseada em aprendizado de máquina para a classificação de elementos em classes

ordenadas, estudar a acurácia da metodologia basando a probabilidade conjunta em distribuições assimétricas e realizar um aprimoramento do pacote CPP-cor.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AGGARWAL, C. C.; REDDY, C. K. **Data Clustering: Algorithms and Applications**. 1st. ed. [S.l.]: Chapman & Hall/CRC, 2013. ISBN 1466558210, 9781466558212.

BASGALUPP, M. porto. **Um algoritmo genérico multi-objetos lexicográficos para indução de árvores de decisão.** Tese (Doutorado) — Instiututo de Ciências Matemáticas e de Computação ICMC - USP, USP, 2 2010.

BELSLEY, D. A.; KUH, E.; WELSCH., R. E. **Regression diagnostics : identifying influential data and sources of collinearity / , Welsch.** New York (N.Y.): Wiley, 1980. Wiley series in probability and mathematical statistics. ISBN 0471058564.

BOUVEYRON, C. et al. **Model-Based Clustering and Classification for Data Science: With Applications in R**. [S.l.]: Cambridge University Press, 2019. (Cambridge Series in Statistical and Probabilistic Mathematics).

BREIMAN, L. Statistical modeling: the two cultures. **Statist. Sci.**, v. 16, n. 3, p. 199–231, 2001. ISSN 0883-4237. With comments and a rejoinder by the author. Disponível em: <http://dx.doi.org/10.1214/ss/1009213726>.

BREIMAN, L. et al. Classification and regression trees chapman & hall. **New York**, 1984.

BREIMAN, L.; FRIEDMAN, J. H. Estimating optimal transformations for multiple regression and correlation. **Journal of the American Statistical Association**, Taylor & Francis, v. 80, n. 391, p. 580–598, 1985. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1985.10478157>.

CAILLAUX, M. A. et al. Container logistics in mercosur: Choice of a transhipment port using the ordinal copeland method, data envelopment analysis and probabilistic composition. **Maritime Economics & amp; Logistics**, v. 13, n. 4, p. 355–370, 2011. Disponível em: <https://EconPapers.repec.org/RePEc:pal:marecl:v:13:y:2011:i:4:p:355-370>.

CANO, J.-R. et al. Monotonic classification: an overview on algorithms, performance measures and data sets. 2018.

CANO, J.-R.; LUENGO, J.; GARCíA, S. Label noise filtering techniques to improve monotonic classification. 2018.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml>.

EVERITT, B. S.; LANDAU, S.; LEESE, M. **Cluster Analysis**. 4th. ed. [S.l.]: Wiley Publishing, 2009. ISBN 0340761199, 9780340761199.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2011.

FERREIRA, D. F. **Estatística Multivariada**. [S.l.]: UFLA, 2018.

FRIZZARINI, C. **Algoritmos para indução de árvores de classificação para dados desbalanceados.** Tese (Doutorado) — USP, São Paulo, 11 2015.

GAVIÃO, L. O. et al. Composição probabilística de preferências com com abordagem empírica em problemas multicritério. **Gestão & produção**, v. 26, 05 2019.

GENTLE, J. E. **Computational Statistics**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2009. ISBN 0387981438, 9780387981437.

GENZ, A. Numerical computation of the multivariate normal probabilities. **The Journal Comput. Graph. Stat.**, p. 141–150, 1992.

GENZ, A. et al. **mvtnorm: Multivariate Normal and t Distributions**. [S.l.], 2020. R package version 1.1-0. Disponível em: <https://CRAN.R-project.org/package=mvtnorm>.

GILLEY, O. W.; PACE, K. On the harrison and rubinfeld data. **Journal of Environmental Economics and Management**, v. 31, n. 3, p. 403–405, 1996. Disponível em: <https://EconPapers.repec.org/RePEc:eee:jeeman:v:31:y:1996:i:3:p:403-405>.

GITHUB, I. **Open Source Survey**. [S.l.]: GitHub, 2019. <https://github.com/github/open-source-survey>.

GUTIÉRREZ, P. A.; GARCÍA, S. Current prospects on ordinal and monotonic classification. **Progress in Artificial Intelligence**, v. 5, n. 3, p. 171–179, August 2016. ISSN 2192-6352. Disponível em: <http://dx.doi.org/10.1007/s13748-016-0088-y>.

HERRERA, F. et al. **Multilabel Classification : Problem Analysis, Metrics and Techniques**. 1. ed. [S.l.]: Springer International Publishing, 2016. ISBN 978-3-319-41110-1,978-3-319-41111-8.

HERRISON, D.; RUBINFELD, D. Hedonic housing prices and the demand for clean air. **Journal of Environmental Economics and Management**, v. 5, n. 1, p. 81–102, 1976.

IMTIAZ, S.; BRIMICOMBE, A. J. A better comparison summary of credit scoring classification. In: . [S.l.: s.n.], 2017.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: <https://faculty.marshall.usc.edu/gareth-james/ISL/>.

KASSAMBARA, A. **Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning**. [S.l.]: STHDA, 2017. v. 1.

KOTŁOWSKI, W.; SŁOWIńSKI, R. On nonparametric ordinal classification with monotonicity constraints. In: . [S.l.: s.n.], 2013.

KRASKER, W. S.; KUH, E.; WELSCH, R. E. **Estimation for Dirty Data and Flawed Models**. North-Holland, Amsterdam: Handbook of Econometrics, 1983. v. 1.

LANGE, N.; RYAN, L. Assessing normality in random effects models. **The Annals of Statistics**, JSTOR, p. 624–642, 1989.

MARSAGLIA, G.; OLKIN, I. Generating correlation matrices. **SIAM Journal on Scientific and Statistical Computing**, v. 5, n. 2, p. 470–475, 1984. Disponível em: <https://doi.org/10.1137/0905034>.

MARTINS, E. F. **Instrumento híbrido aplicado ao estudo da confiabilidade humana em evento de perda de energia elétrica externa em usina nuclear**. Tese (Doutorado) — Universidade Federal Fluminense, 2015.

MICHAEL, D.; CONSTANTIN, Z. **Multicriteria decision aid classification methods**. 2nd. ed. [S.l.]: Springer US, 2002. ISBN 9780306481055.

MORETTIN, P. **Estatística básica**. [S.l.]: Saraiva Educação S.A., 2017. ISBN 9788547220235.

NETO, J. M. et al. **Estatística multivariada: uma visão didática-metodológica**. 2015.

OLIVEIRAL DIMAS SAMID LEME, B. H. G. B. M. P. R. R. G. F. A. P. Emanuelle Morais de. A computer vision system for coffee beans classification based on computational intelligence techniques. **Journal of Food Engineering**, 2016.

PACE, R. K. Nonparametric methods with applications to hedonic models. **The Journal of Real Estate Finance and Economics**, Springer, v. 7, n. 3, p. 185–204, 1993.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

REQUENA, G. de L. **Predições estatísticas para dados politômicos**. 62 p. Tese (Doutorado), 2018.

RIBEIRO, M. et al. **cppcor: Probabilistic Composition of Correlated Preference**. [S.l.], 2020. R package version 1.2-0. Disponível em: <https://cran.r-project.org/web/packages/cppcor/index.html>.

ROY, B.; SKALKA, J. **ELECTRE IS : Aspects méthodologiques et guide d'utilisation**. [S.l.], 1984. 125 pp. p.

SANT'ANNA, A. P. Aleatorização e composição de medidas de preferências. **Pesquisa Operacional**, v. 22, n. 1, p. 87 – 103, 2002.

SANT'ANNA, A. P. Detalhamento de uma metodologia de classificação baseada na composição probabilística de preferências. **RELATÓRIOS DE PESQUISA EM ENGENHARIA DE PRODUÇÃO**, v. 13, n. 2, p. 12–21, 2013. Disponível em: <http://www.producao.uff.br/conteudo/rpep/volume132013/RelPesq_V13_2013_C02.pdf>.

SANT'ANNA, A. P. Detalhamento de uma Metodologia de Classificação baseada na Composição Probabilítica de Preferências. **RPEP**, v. 13, n. 2, p. 9, 2013. ISSN 1678 - 2399.

SANT'ANNA, A. P. Procedimento de Cálculo para a Composição Probabilítica de Preferências. **RPEP**, v. 13, n. 1, p. 9, 2013. ISSN 1678-2399.

SANT'ANNA, A. P. **Probabilistic composition of preferences, theory and applications**. [S.l.]: SPRINGER INTERNATIONAL PU, 2015.

SANT'ANNA, A. P. **Probabilistic Composition of Preferences, Theory and Applications**. 2015.

SANT'ANNA, A. P.; CONDE, F. Q. Probabilistic comparison of call centres in a group decision process. **International Journal of Management and Decision Making**, v. 11, n. 5/6, p. 417–437, 2011. Disponível em: <https://ideas.repec.org/a/ids/ijmdma/v11y2011i5-6p417-437.html>.

SANT'ANNA, A. P.; COSTA, H. G.; PEREIRA, V. Cpp-tri: um método de classificação ordenada baseado em composição probabilística. **Relatórios de Pesquisa em Engenharia de Produção**, v. 12, n. 8, p. 104–117, 2012.

SANT'ANNA, A. P.; COSTA, H. G.; PEREIRA, V. CPP-TRI: a sorting method based on the probabilistic composition of preferences. **International Journal of Information and Decision Sciences**, Inderscience Publishers, v. 7, n. 3, p. 193, 2015.

SANT'ANNA, A. P.; FARIA, F.; COSTA, H. G. APLICAção DA COMPOSIção PROBABILíSTICA e DO méTODO DAS k-méDIAS à CLASSIFICAção DE MUNICíPIOS QUANTO à OFERTA DE CRECHES. **Cadernos do IME - Série Estatística**, Universidade de Estado do Rio de Janeiro, v. 34, n. 1, jun 2013.

SANT'ANNA, A. P.; FERREIRA, M.; DUARTE, S. dos R. A. AvaliaÇão do desempenho de empresas utilizando a composiÇão probabilística de índices financeiros. In: . [S.l.: s.n.], 2012.

SANTANNA, A. P. et al. Beta Distributed Preferences in the Comparison of Failure Modes. **Procedia Computer Science**, v. 55, p. 862–869, 2015. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050915016245>.

SANT'ANNA, A. P.; MELLO, J. A. C. C. B. S. d. Validating rankings in soccer championships. **Pesquisa Operacional**, scielo, v. 32, p. 407 – 422, 08 2012. ISSN 0101-7438. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382012000200008&nrm=iso>.

SANT'ANNA, A. P.; SANT'ANNA, L. A. F. P. Randomization as a stage in criteria combining. In: **VII International Conference on Industrial Engineering and Operations Managerment, Salvador, (ICIEOM)**. [S.l.: s.n.], 2001. p. 248–256.

SANT'ANNA, A. Probabilistic human development indices. **Brazilian Journal of Operations & Production Management**, v. 12, n. 1, p. 136–146, Jun. 2015. Disponível em: <https://bjopm.emnuvens.com.br/bjopm/article/view/V12N1A13>.

SHAHHOSSEINI, M.; HU, G.; PHAM, H. A case studyfor housing price prediction. **Industrial and Manufacturing Systems Engineering Conference Proceedings and Posters**, p. 185, 2019.

SHANKAR, S.; CARSON, R. T. **Robust Regression in the Presence of Heteroskedasticity**. [S.l.]: JAI Press, 1988. v. 7. (Advances in Econometrics, v. 7).

SHARAFF, A.; NAGWANI, N. K.; DHADSE, A. Comparative study of classification algorithms for spam email detection. In: . [S.l.: s.n.], 2016.

SILVA, G. B. da. **Combining data mining techniques with multicriteria decision aid in classification problems with composition probabilistc of preferences in trichotomic procedure (CPP-TRI)**. 164 p. Tese (Doutorado), 2016.

SILVA, G. B. da; BARROS, M. D. de; COSTA, H. G. ABORDAGEM híBRIDA MULTICRITéRIO – MINERAção DE DADOS APLICADA a CLASSIFICAção DE UNIDADES DA FEDERAção COM BASE NA POPULAção ECONOMICAMENTE OCUPADA. In: **Anais do XVIII Simpósio de Pesquisa Operacional & Logística da Marinha**. [S.l.]: Editora Edgard Blücher, 2016.

SPRENT, P.; SMEETON, N. C. **Applied nonparametric statistical methods**. [S.l.]: Chapman and Hall/CRC, 2000.

SUZUKI, E. et al. Gene expression profile of peripheral blood mononuclear cells may contribute to the identification and immunological classification of breast cancer patients. **Breast Cancer**, v. 26, n. 3, p. 282–289, May 2019. ISSN 1880-4233. Disponível em: <https://doi.org/10.1007/s12282-018-0920-2>.

TABAK, M. A. et al. Machine learning to classify animal species in camera trap images: applications in ecology. **bioRxiv**, Cold Spring Harbor Laboratory, 2018. Disponível em: <https://www.biorxiv.org/content/early/2018/07/09/346809>.

TAN, P.-N. et al. **Introduction to Data Mining (2Nd Edition)**. 2nd. ed. [S.l.]: Pearson, 2018. ISBN 0133128903, 9780133128901.

THEODORIDIS, S. **Introduction to pattern recognition: a MATLAB approach**. [S.l.]: Elsevier/Academic, 2010.

TREINTA, F. T. et al. Metodologia de pesquisa bibliográfica com a utilizaçã de método multicritério de apoio à decisã. **Production**, v. 24, n. 2, p. 508–520, 09 2014.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. Berlin, Heidelberg: Springer-Verlag, 1995. ISBN 0-387-94559-8.

WILCOXON, F. Individual comparisons by ranking methods. In: . [S.l.: s.n.], 1945.

WILLI, M. et al. Identifying animal species in camera trap images using deep learning and citizen science. In: . [S.l.: s.n.], 2019.

YEVSEYEVA, I. **Solving classification problems with multicriteria decision aiding approaches**. [S.l.]: University of Jyväskylä, 2007.

ZOPOUNIDIS, C.; DOUMPOS, M. Multicriteria classification and sorting methods: A literature review. **European Journal of Operational Research**, v. 138, n. 2, p. 229–246, April 2002. Disponível em: <https://ideas.repec.org/a/eee/ejores/v138y2002i2p229-246.html>.