UNIVERSIDADE PRESBITERIANA MACKENZIE

RENÊ DE ÁVILA MENDES

Composição de um Indicador de Qualidade para Classificações Binárias com Base na Qualidade e na Complexidade dos Dados

São Paulo SP

2021

RENÊ DE ÁVILA MENDES

Composição de um Indicador de Qualidade para Classificações Binárias com Base na Qualidade e na Complexidade dos Dados

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Presbiteriana Mackenzie como requisito para a obtenção do título de Doutor em Engenharia Elétrica e Computação, área de concentração Engenharia da Computação.

Orientador: Prof. Dr. Leandro Augusto da Silva

São Paulo SP

2021

Elaborado pelo Sistema de Geração Automática de Ficha Catalográfica da Mackenzie
com os dados fornecidos pelo(a) autor(a)

M538c Mendes, Renê de ávila

Composição de um Indicador de Qualidade para Classificações Binárias com Base na Qualidade e na Complexidade dos Dados: [recurso eletrônico] / Renê de ávila - Mendes.

2897 KB ; il.

Tese (Doutorado em Engenharia Elétrica e Computação) - Universidade Presbiteriana Mackenzie, São Paulo, 2022.

Orientador(a): Prof(a). Dr(a). Leandro Augusto da Silva

Referências Bibliográficas: f. 103 -108

1. Complexidade de Dados. 2. Qualidade de Dados. 3. Classificação de Dados. 4. sem. 5. Pls-sem. I. Silva, Leandro Augusto da, *orientador(a)*.II. Título.

Bibliotecário Responsável: Maria Gabriela Brandi Teixeira - CRB 8/6339

Folha de Identificação da Agência de Financiamento

Autor: Renê de Ávila Mendes

Programa de Pós-Graduação *Stricto Sensu* em Engenharia Elétrica e Computação

Título do Trabalho: Composição de um Indicador de Qualidade para Classificações Binárias com Base na Qualidade e na Complexidade dos Dados

O presente trabalho foi realizado com o apoio de [1]:

☐ CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

☐ CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

☐ FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo

☒ Instituto Presbiteriano Mackenzie/Isenção integral de Mensalidades e Taxas

☒ MACKPESQUISA - Fundo Mackenzie de Pesquisa

☐ Empresa/Indústria:

☐ Outro:

[1] **Observação**: caso tenha usufruído mais de um apoio ou benefício, selecione-os.

RENÊ DE ÁVILA MENDES

COMPOSIÇÃO DE UM INDICADOR DE QUALIDADE PARA CLASSIFICAÇÕES BINÁRIAS COM BASE NA QUALIDADE E NA COMPLEXIDADE DOS DADOS

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Universidade Presbiteriana Mackenzie como requisito para a obtenção do título de Doutor em Engenharia Elétrica e Computação, área de concentração Engenharia da Computação.

Aprovada em 10/12/2021

BANCA EXAMINADORA

Leandro A. Silva

Prof. Dr. Leandro Augusto da Silva
Universidade Presbiteriana Mackenzie

Prof. Dr. Diógenes de Souza Bido
Universidade Presbiteriana Mackenzie

Prof. Dr. Arnaldo Rabello de A. Vallim Filho
Universidade Presbiteriana Mackenzie

Prof. Dr. André Ponce de Leon F. de Carvalho
Universidade de São Paulo

Profa. Dra. Alexandra Aparecida de Souza
Instituto Federal de São Paulo

A DEUS, para Sua glória.
À KATYA, minha amada esposa.
À SARA e à ESTER, minhas
preciosas filhas.

Agradecimentos

Essa pesquisa, seus resultados, a conquista do título de Doutor e as implicações dessa conquista são concessões das bondosas e amorosas mãos do **DEUS** triúno, criador e sustentador do Universo. Assim, minha humilde, sincera e emocionada gratidão à Pessoa bendita do **DEUS Pai, ao Seu Filho, JESUS, o DEUS encarnado, e à Pessoa bendita do ESPÍRITO SANTO**.

À minha amada Esposa **KATYA** minha gratidão especial, por ter acreditado e investido em mim. Às minhas preciosas filhas, **SARA e ESTER**, meu carinho e minha gratidão pelo apoio e pela paciência comigo.

Ao meu amado Pai, **Prof. Dr. MARCEL MENDES**, minha gratidão pelo senhor ser esse cristão honrado, servo dedicado ao Reino de DEUS, homem não apenas inteligente e culto, mas sábio.

Agradecimentos especiais são direcionados à **UNIVERSIDADE PRESBITERIANA MACKENZIE**, que proporcionou bolsa integral para a realização destes estudos, ao **Fundo MACKPESQUISA** e ao **Laboratório BigMAAp - Big Data e Métodos Analíticos Aplicados**.

Ao **Prof. Dr. LEANDRO AUGUSTO DA SILVA** minha gratidão por ter pacientemente me orientado nas pesquisas e pelo apoio para as publicações.

Ao **INSTITUTO PRESBITERIANO MACKENZIE**, nas pessoas do senhores **ADILSON PEREIRA e RICARDO FUKUDA MARQUES**, minha gratidão pelo apoio que recebi na difícil tarefa de conciliar as jornadas de trabalho e de estudo.

Aos Professores membros da Banca Examinadora, **Prof. Dr. DIÓGENES DE SOUZA BIDO, Prof. Dr. ARNALDO RABELLO DE A. VALLIM FILHO, Prof. Dr. ANDRÉ PONCE DE LEON FERREIRA DE CARVALHO e Profa. Dra. ALEXANDRA APARECIDA DE SOUZA**, minha gratidão pelas preciosas contribuições para este trabalho.

"[...] a fim de conhecerem plenamente o mistério de Deus,
a saber, Cristo.
Nele estão escondidos todos os tesouros da sabedoria e do conhecimento."
(Bíblia Sagrada, Colossenses capítulo 2, versos 2 e 3)

Resumo

A classificação de dados é uma tarefa de mineração de dados que consiste na aplicação de um algoritmo a conjunto de dados de treinamento com a finalidade de inferir a classe de um objeto (não classificado) em análise. Uma parte significativa do desempenho do algoritmo de classificação depende da complexidade e da qualidade do conjunto de dados. A Complexidade dos Dados envolve a investigação dos efeitos da dimensionalidade, da sobreposição de atributos e da separabilidade das classes. A Qualidade dos Dados, no que lhe concerne, se concentra em aspectos como ruídos e valores ausentes. Na literatura são poucos os estudos que debatem a relação entre os fatores, complexidade e qualidade, visando ponderar a influência de cada um na qualidade do desempenho de um algoritmo. Esta pesquisa aplica a Modelagem de Equações Estruturais (SEM) e o algoritmo *Partial Least Squares Structural Equation Modeling* (PLS-SEM) e, de forma inovadora, apresenta um indicador composto, chamado de Indicador de Qualidade de Classificação para conjuntos de dados binários (IQCb), que associa as contribuições da Complexidade dos Dados e da Qualidade dos Dados para a Qualidade da Classificação. A modelagem experimental com 178 conjuntos de dados obtidos do repositório OpenML mostrou que o controle da complexidade melhora os resultados da classificação mais do que a qualidade dos dados. Adicionalmente, esta tese também apresenta uma ferramenta visual para a avaliação de conjuntos de dados quanto ao desempenho de classificação.

Palavras-chaves: Complexidade de Dados. Qualidade de Dados. Classificação de Dados. SEM. PLS-SEM.

Abstract

Data classification is a data mining task that consists in applying an algorithm to a training dataset in order to infer the class of an (unclassified) object under analysis. A significant part of the classification algorithm's performance depends on the dataset's complexity and quality. Data Complexity involves the investigation of the effects of dimensionality, the overlap of descriptive attributes, and the classes' separability. Data Quality, as far as it is concerned, focuses on aspects such as noise and missing values. There are few studies in the literature that discuss the relationship between complexity and quality aiming to consider the influence of each on the quality of an algorithm's performance. This research applies Structural Equation Modeling (SEM) and the *Partial Least Squares Structural Equation Modeling* (PLS-SEM) algorithm and, in an innovative way, presents a composite indicator, called Classification Quality Indicator for sets of binary data (IQCb), which associates the contributions of Data Complexity and Data Quality to Classification Quality. Experimental modeling with 178 datasets obtained from the OpenML repository showed that controlling complexity improves classification results more than data quality. Additionally, this thesis also presents a visual tool for evaluating datasets for classification performance.

Keywords: Data Complexity. Data Quality. Data Classification. SEM. PLS-SEM.

Lista de Figuras

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under the ROC Curve*
CART	*Classification And Regression Trees*
CQ	*Classification Quality* (Qualidade da Classificação)
DC	*Data Complexity* (Complexidade de Dados)
DQ	*Data Quality* (Qualidade de Dados)
IQCb	Indicador de Qualidade de Classificação para bases de dados binárias
KDD	*Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bases de Dados)
MD	*Data Mining* (Mineração de Dados)
PLS-SEM	*Partial Least Squares Structural Equation Modeling*
RF	*Random Forests*
SEM	*Structural Equation Modeling*

Sumário

1 Introdução

O crescimento da quantidade de dados disponível e a necessidade de métodos e técnicas que pudessem transformar esses dados em conhecimento se tornou a preocupação de uma área chamada Descoberta de Conhecimento em Bases de Dados (KDD - *Knowledge Discovery in Databases*). O processo de descoberta de padrões proposto pelo KDD prevê a preparação, a seleção e a limpeza dos dados como predecessores obrigatórios da Mineração de Dados (DM) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), demonstrando a preocupação do processo de KDD com a qualidade dos resultados produzidos.

Na busca por seu objetivo principal de encontrar padrões de dados, a área de KDD invade as fronteiras e é beneficiada por outras áreas do conhecimento, tais como Inteligência Artificial, Aprendizado de Máquina, Visualização de Dados, Computação de Alto Desempenho e outras áreas cujo objetivo seja o de extrair conhecimento de alto nível a partir de dados de baixo nível no contexto de grandes conjuntos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; FAYYAD, 1997).

Mais recentemente, a profusão de sensores, de dispositivos conectados à *web*, de interações pessoais com aplicações *mobile* e da ciência computacional intensiva (HEY et al., 2009) apresentou-se como um desafio para o processamento, a análise e a compreensão de dados, com algumas dimensões distintivas de problemas anteriormente tratados pela KDD: a variedade e a velocidade dos dados. O termo *Big Data* foi atribuído aos problemas que combinam essas duas dimensões dos dados ultimamente percebidas àquelas já identificadas anteriormente por Fayyad, Piatetsky-Shapiro e Smyth (1996): volume e variedade. Essa nomenclatura foi cunhada por John Mashey, Cientista Chefe da empresa Silicon Graphics, em 1998, em seu trabalho intitulado *"Big Data and the Next Wave of InfraStress"*, e utilizada academicamente em seu sentido atual inicialmente por Sholom M. Weiss e Nitin Indurkhya, em 1998, em seu livro *"Predictive Data Mining: A Practical Guide"*, e em 2003 por Francis X. Diebold em sua publicação intitulada *"'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting"* (DIEBOLD, 2012).

A intersecção entre as áreas de conhecimento KDD e *Big Data* parece estender-se além das dimensões dos dados para a sua multidisciplinaridade e a sua aplicação a áreas de pesquisa tais como Bioquímica, Astronomia, Bio-Informática, Economia, Negócios e Sociologia (KAMBATLA et al., 2014; CHEN; ZHANG, 2014; FANG et al., 2015). A variedade de origens, padrões e formatos de dados que compõe o *Big Data* ao mesmo tempo que confere a riqueza de onde será extraída a informação de valor também incorpora aos processos analíticos a preocupação com a qualidade dos resultados (CHEN; ZHANG,

2014; ANAGNOSTOPOULOS; ZEADALLY; EXPOSITO, 2016). Assim, os passos que precedem a Mineração de Dados no KDD não se dispensam em contextos *Big Data*.

O processo de KDD impõe cuidados que se expressam na forma de etapas ou passos encadeados e dependentes. Inicia-se com a compreensão do domínio da aplicação e o estabelecimento do objetivo do processo de KDD, seguidos da aquisição dos dados, objetos da análise. Os dados são então preparados para a análise, o que pode incluir operações de limpeza, integração, redução, transformação e discretização, e cujo resultado torna o conjunto de dados apto à aplicação de métodos de mineração de dados compatíveis com os objetivos estabelecidos no primeiro passo do processo de KDD. Preparados os dados, passa-se à sua análise exploratória e à escolha dos algoritmos, atividades essenciais para a Mineração dos Dados, que aplica os algoritmos selecionados à busca por um padrão que possa ser representado em um ou mais modelos. Esses modelos descobertos são finalmente interpretados, com o apoio de ferramentas de visualização, podendo essa atividade conduzir a um passo anterior do processo ou ao passo final de incorporação do conhecimento descoberto (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017).

Embora custosa, a obediência aos passos do processo de KDD, que podem ser seguidos de forma iterativa e exigir *loops*, é recompensada pela possibilidade da descoberta de conhecimento. Por outro lado, a negligência de um ou mais passos pode expor os algoritmos de mineração à presença de dados ruidosos, incompletos, inconsistentes, com atributos pouco informativos ou redundantes, ou levar ao sobreajuste do modelo aos dados dos quais foi induzido (*overfitting*).

Assim, o posicionamento das atividades de pré-processamento e transformação como predecessoras essenciais para a mineração de dados e a preocupação de *Big Data* com a dimensionalidade e o volume dos dados são indicativos da influência que o conteúdo e a estrutura dos conjuntos de dados exercem sobre o resultado de uma análise no processo de KDD. Mais especificamente, nota-se na literatura preocupação com seus aspectos internos, tais como distribuição dos dados, ausência de medições e a presença de ruídos e inconsistências, e com seus aspectos estruturais, tais como quantidade de objetos e a dimensionalidade do conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HO; BASU, 2000; LAROSE, 2015; SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017; TAN; STEINBACH; KUMAR, 2018).

Os esforços para definir e quantificar a qualidade do conteúdo dos dados remontam da perspectiva semiótica, que vê os dados como representações de fatos, eventos ou objetos, e passam pela identificação e mensuração de dimensões da qualidade nos dados, que têm sido contabilizadas em mais de uma centena de dimensões (BERTI-EQUILLE, 2007; JAYAWARDENE; SADIQ; INDULSKA, 2015). A Estatística desempenha papel fundamental mensurando e descrevendo as características estruturais e internas de con-

juntos de dados e a literatura parece indicar a organização de uma área de conhecimento especializada em estudar o quanto aspectos como dimensionalidade, esparsamento, resolução, tamanho e, até mesmo, qualidade exercem influência sobre as análises. Mesmo que não formalmente definida, essa área é responsável pelo estudo da complexidade dos dados (HO; BASU, 2000; HO; BASU, 2002; SÁNCHEZ; MOLLINEDA; SOTOCA, 2007; GARCIA; CARVALHO; LORENA, 2015; ZUBEK; PLEWCZYNSKI, 2016; BARELLA et al., 2018; GARCIA; LORENA; LEHMANN, 2018).

1.1 Motivação, Hipótese e Objetivo

Embora haja na literatura a fundamentação teórica para afirmar que atributos de qualidade e atributos de complexidade dos conjuntos de dados exercem influência sobre o resultado das análises, e embora tenham sido definidas métricas que permitam a mensuração de alguns desses atributos, nota-se uma lacuna na discussão de seus efeitos combinados sobre o poder de generalização de modelos da Mineração de Dados.

O uso de medidas objetivas como instrumento de avaliação da descoberta de conhecimento em bases de dados foi destacado por Fayyad, Piatetsky-Shapiro e Smyth (1996) e pouco depois por Ho e Basu (2000), tendo esses últimos notado a relação entre a complexidade estrutural de dados reais e o desempenho de classificadores, e registrado a utilidade da taxa de erro no estudo da complexidade de dados.

A busca pela compreensão da influência combinada da complexidade e da qualidade dos dados sobre os resultados da tarefa de Classificação pode ser entendida como a pesquisa por um modelo que relacione variáveis, ou como um problema de otimização de variáveis. Embora haja instrumentos matemáticos dentro da própria Mineração de Dados que permitam a busca por uma função ótima dentro de um conjunto viável, a exemplo do que podem fazer as redes neurais, nota-se que variáveis como Complexidade de Dados e Qualidade de Dados carecem de uma medida direta que permita sua quantificação. Tal conclusão se deduz da ausência de definição formal para o que se denomina como complexidade de dados e das múltiplas perspectivas sobre o que seja e como se meça a qualidade dos dados (JAYAWARDENE; SADIQ; INDULSKA, 2015).

Nesse contexto, a Modelagem de Equações Estruturais (SEM - *Structural Equation Modeling*) se apresenta como uma ferramenta valiosa: permite discussões de naturezas exploratória ou confirmatória sobre as interações de variáveis, oferecendo subsídios para compreender como as variáveis são construídas e medidas. Além de contribuir com recursos para a construção de um modelo que represente a relação entre variáveis, a Modelagem de Equações Estruturais dispõe de instrumentos de dedução da interação entre as variáveis e entre as variáveis e seus indicadores.

Dentre esses instrumentos destaca-se o método estatístico PLS-SEM *Partial Least*

Squares Structural Equation Modeling, amplamente utilizado na análise exploratória multivariada de dados nas Ciências Sociais (TENENHAUS et al., 2005; ZWICKER; SOUZA; BIDO, 2008). O PLS-SEM usa dados amostrais para estimar a contribuição de variáveis e a força de seus relacionamentos em um modelo SEM, buscando minimizar a variância residual não explicada de variáveis dependentes (HENSELER; RINGLE; SARSTEDT, 2012; HAIR et al., 2016).

É nesse cenário que a presente pesquisa busca se desenvolver, fixando como objetivo a proposta de um indicador composto de qualidade para classificações binárias (apenas duas classes) em conjuntos de dados, tomando por base indicadores de complexidade geométrica e de qualidade obtidos desses conjuntos de dados. A pesquisa pela literatura encontrou metodologias de análise da qualidade de conjuntos de dados baseadas na análise descritiva dos dados (SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017), ou apenas na análise de dimensões da qualidade dos dados (JAYAWARDENE; SADIQ; INDULSKA, 2015) ou por metodologias visuais (TENG et al., 2012), mas não encontrou precedentes de elaboração de um indicador composto que relacione a complexidade e a qualidade dos dados por meio de um modelo SEM.

No contexto da Análise Preditiva, especificamente na tarefa de Classificação, um indicador composto de qualidade de classificação poderia oferecer maior compreensão sobre a influência combinada que a complexidade e a qualidade dos dados exercem sobre o desempenho de classificadores. O uso do indicador composto de qualidade poderia ainda contribuir para o aumento da qualidade das análises, ou mesmo permitir comparações de desempenho entre diferentes conjuntos de dados e entre diferentes classificadores, ou ainda permitir a previsão do desempenho de classificadores com base em características intrínsecas das amostras.

Como objetivo secundário a presente pesquisa espera apresentar um método visual de comparação de conjuntos de dados tomando por base suas dimensões de qualidade e complexidade.

A afirmação de que é possível estimar o desempenho da tarefa de classificação binária de conjuntos de dados reais com base em atributos de qualidade e de complexidade do conjunto de dados analisado se expressará na pesquisa por meio da seguinte hipótese central: quanto maior o valor do indicador de qualidade de um conjunto de dados binário real melhores os resultados de classificadores binários.

1.2 Organização do Documento

O documento de pesquisa se organiza da seguinte maneira: o Capítulo 2 apresenta a revisão da literatura e define o problema; o Capítulo 3 explora a metodologia de Equações Estruturais e as dimensões de complexidade e qualidade que afetam a tarefa de

classificação de dados; o Capítulo 4 apresenta os modelos propostos na pesquisa e descreve os procedimentos metodológicos que relacionam as dimensões identificadas aos resultados da classificação de dados; o Capítulo 5 apresenta e discute os resultados e a aplicabilidade; finalmente, o Capítulo 6 conclui a pesquisa apresentando suas contribuições, limitações e as oportunidades de pesquisas futuras.

2 Trabalhos Correlatos

Algumas das referências acadêmicas importantes sobre Qualidade de Dados (DQ) datam da década de 1990, da perspectiva semiótica dos dados como representação de fatos, objetos ou pessoas (LIEBENAU; BACKHOUSE, 1990), ou da perspectiva declarativa, que vê os dados como matéria-prima para informações (WANG; STRONG, 1996). Na perspectiva declarativa, as dimensões de qualidade intrínseca que explicam os dados podem ser agrupadas, como aquelas impostas pelos metadados, padrões de esquema ou regras de negócios. Do ponto de vista do uso existem dimensões cuja avaliação depende do usuário, como as relacionadas à eficiência e eficácia da criação e usabilidade dos dados. As dimensões podem ser classificadas em termos da granularidade em que se aplicam: a um elemento de dados (atributo de uma entidade), a um registro de dados (coleção de recursos que constituem uma entidade) ou a um objeto de informação (coleção de registros) (JAYAWARDENE; SADIQ; INDULSKA, 2015).

A aplicação de DQ na análise de dados encontra diferentes importâncias para as dimensões. Por exemplo, a pesquisa de Berti-Equille (2007) mostra o efeito das variações de atualidade, acurácia, completude e consistência na mineração de regras de associação. Por outro lado, Hair et al. (2014) apontam valores discrepantes e valores ausentes como problemas mais expressivos na Análise Multivariada de Dados, descrevendo os procedimentos de identificação e abordagem desses problemas de qualidade em conjuntos de dados. Estudos recentes relataram os efeitos de problemas de qualidade de dados em processos de migração de dados (JANUZAJ; JANUZAJ, 2009; AZEROUAL; JHA, 2021), sistemas de mineração de dados (JANUZAJ; JANUZAJ, 2009), sistemas de Big Data Analytics (WOOK et al., 2021; TALEB et al., 2021), sistemas de Internet das Coisas (IoT) (KARKOUCH et al., 2016) e em Engenharia de Software (ROSLI; TEMPERO; LUXTON-REILLY, 2013; VALVERDE et al., 2014; BOSU; MACDONELL, 2019). Além disso, as dimensões dos dados identificadas como relevantes variam de acordo com o foco do estudo (LARANJEIRO; SOYDEMIR; BERNARDINO, 2015).

O método de Modelagem de Equações Estruturais (SEM) e o algoritmo *Partial Least Squares Structural Equation Modeling* (PLS-SEM) são aplicados por Azeroual e Jha (2021) para analisar a relação entre o sucesso de uma migração de dados entre sistemas e problemas de qualidade de dados, especificamente correção, completude, consistência e atualidade (*timeliness*), e também por Wook et al. (2021) para medir o efeito das características de Big Data (os vários Vs de Big Data) em aspectos de qualidade de dados que podem afetar a análise de dados em Big Data. Algumas das dimensões de qualidade de dados consideradas relevantes para análise são acurácia, credibilidade, completude, atualidade e facilidade de operação (WOOK et al., 2021). Na linha de pesquisa de qualidade

de dados para sistemas de Big Data, Taleb et al. (2021) propoem o *Big Data Quality Management Framework* (BDQMF) para identificar e solucionar problemas de qualidade de dados do ciclo de vida de Big Data. Diversas dimensões da qualidade dos dados são identificadas e tratadas no BDQMF, agrupadas em dimensões intrínsecas (completude, consistência, acurácia, atualidade), dimensões contextuais (credibilidade, relevância, valor agregado, quantidade, acessibilidade, reputação), dimensões de acessibilidade (acesso, segurança) e dimensões representacionais (interpretabilidade, manipulabilidade, facilidade de compreensão, concisão da representação, consistência representacional). No sentido de identificar e quantificar problemas de qualidade de dados, técnicas de Mineração de Dados, tais como Agrupamento, Agrupamento de Subspaços (*Subspace Clustering*) e Classificação de Dados, foram utilizadas por Januzaj e Januzaj (2009), e a técnica de teste metamórfico, utilizada em testes de *software*, foi proposta por Auer e Felderer (2019).

A Classificação de Dados é uma tarefa de Mineração de Dados que aplica um algoritmo a objetos de um conjunto de dados de treinamento e um atributo classificatório para prever a classe de um objeto (não classificado) na análise. Mais recentemente, aspectos estruturais dos dados também foram identificados como relevantes na classificação dos dados, possivelmente devido à sensibilidade desta tarefa às características geométricas dos dados. Esses aspectos estruturais dos dados foram estudados sob o nome de Complexidade de Dados (DC). Tipos de desafios para tarefas de classificação relacionadas a DC são identificados: a) a ambiguidade de classes, que ocorre quando os atributos do conjunto de dados não são suficientes para um algoritmo de classificação distinguir entre as classes, ou as classes não estão claramente definidas ou os atributos não são informativos o suficiente para separação de classes; b) complexidade da fronteira, cujo grau pode ser medido pela quantidade de informação necessária para descrever o limite entre as classes em um conjunto de dados; e c) esparsidade da amostra e dimensionalidade de espaço de atributos, que ocorre quando a capacidade de generalização de um classificador é prejudicada por amostras que podem representar conjuntos de dados insuficientemente, mesmo quando o espaço de características é grande, aumentando a variabilidade da área de decisão do classificador (HO; BASU, 2002; HO; BASU; LAW, 2006).

A preocupação com o desempenho de algoritmos de classificação relacionados a DQ e DC é compartilhada na literatura por outros pesquisadores, como Sánchez, Mollineda e Sotoca (2007), que discutem o efeito negativo de algumas dimensões da complexidade de dados no algoritmo de aprendizagem supervisionada k vizinhos mais próximos (k-NN), como alta dimensionalidade de dados, sobreposição de classes e densidade. A presença de ruídos no componente de classe de um conjunto de dados, usando medidas de sobreposição e separabilidade de classes, geometria e topologia, e representação estrutural de dados é explorada por Garcia, Carvalho e Lorena (2015).

A relação entre as dimensões da DC e o desempenho dos algoritmos de classifica-

ção é abordada posteriormente por Barella et al. (2018). Este estudo aplica medidas de complexidade em conjuntos de dados reais e artificiais para identificar o efeito da sobreposição de classes em tarefas de classificação onde as classes são desbalanceadas. A busca pela visualização da relação entre DC e seus efeitos na análise dos dados é abordada por Zubek e Plewczynski (2016), que apresentam uma medida de complexidade visual com aplicação direta na redução de dados do estágio de treinamento de classificadores.

Na revisão da literatura verificou-se a ausência de um estudo dos efeitos combinados de DQ e DC sobre o desempenho de algoritmos de classificação, ou Qualidade de Classificação (CQ). Um resultado de pesquisa altamente relevante é encontrado em Blake e Mangiameli (2011), os quais analisaram quatro problemas de DQ (acurácia, completude, consistência e atualidade) e um problema de DC (entropia) introduzido artificialmente para testar seus efeitos no CQ. A medida-F de seis algoritmos de classificação de dados (Multilayer Perceptron (MLP), J48, Otimização mínima sequencial (SMO), IBk, rede bayesiana e regressão logística) mediu a qualidade da classificação de dados.

A construção de indicadores compostos é explorada com detalhes por Nardo et al. (2008), que propoem e detalham uma metodologia de construção, desde a fundamentação teórica, passando pela escolha das variáveis, dos métodos de tratamento de dados ausentes, análise dos dados, ponderação e agregação dos indicadores, até os testes de robustez do indicador composto obtido. Uma atualização da pesquisa de ponderação e agregação dos indicadores individuais e dos testes de robustez é apresentada por Greco et al. (2018). Usando a Modelagem de Equações Estruturais e o algoritmo PLS-SEM Libório et al. (2020) propõem um novo indicador composto de desigualdade intra-urbana, enquanto Tomaselli, Fordellone e Vichi (2020) combinam os algoritmos PLS-SEM e k-NN para construir um indicador composto de bem-estar em microterritórios.

A presente pesquisa tenta preencher esta lacuna: estudar o efeito combinado de DQ e DC no QC, procurando na literatura os indicadores mais recentes para essas variáveis e uma metodologia que quantifique essa relação.

3 Uma Introdução a Modelagem de Equações Estruturais

Neste capítulo são apresentados os conceitos necessários ao desenvolvimento da pesquisa, incluindo Modelagem de Equações Estruturais, PLS-SEM, Qualidade de Dados e Complexidade de Dados.

3.1 Modelagem de Equações Estruturais e PLS-SEM

Partial Least Squares Structural Equation Modeling (PLS-SEM), também conhecido como PLS *path modeling*, é um método estatístico de segunda geração aplicado à análise exploratória multivariada de dados. Essa técnica é considerada *soft model basic design* por contrastar com a *Covariance-Based Structural Equation Modeling* (CB-SEM), cujas suposições quanto ao modelo de equações estruturais, à distribuição dos dados e ao tamanho da amostra são mais restritivas. Quanto ao uso, PLS-SEM é aplicado principalmente no desenvolvimento de teorias na análise exploratória, diferindo do uso primariamente confirmatório do método CB-SEM.

O procedimento de estimação do PLS-SEM baseia-se no método dos mínimos quadrados (OLS) baseado em regressão, que usa os dados para estimar os coeficientes dos relacionamentos do modelo na Modelagem de Equações Estruturais (SEM - *Structural Equation Modeling*). O objetivo do método PLS-SEM é minimizar a variância residual não explicada dos construtos endógenos (construtos que estão sendo explicados por outros construtos no modelo), buscando coeficientes para os relacionamentos que maximizem os valores R^2 dos construtos endógenos (TENENHAUS et al., 2005; HAIR et al., 2016; SARSTEDT; RINGLE; HAIR, 2017).

Nesta seção são detalhados os elementos do método PLS-SEM que fundamentarão a pesquisa.

3.1.1 Modelagem de Equações Estruturais

A estrutura fundamental da Modelagem de Equações Estruturais (SEM) é o modelo (do inglês *path model*), diagrama que representa os relacionamentos entre variáveis e as hipóteses que fundamentam os relacionamentos e a distribuição das variáveis. O modelo é o que se avalia ao aplicar a SEM.

Mais detalhadamente, o modelo da SEM (Figura 1) é construído pela disposição de construtos, de indicadores, dos relacionamentos entre os construtos e seus indicado-

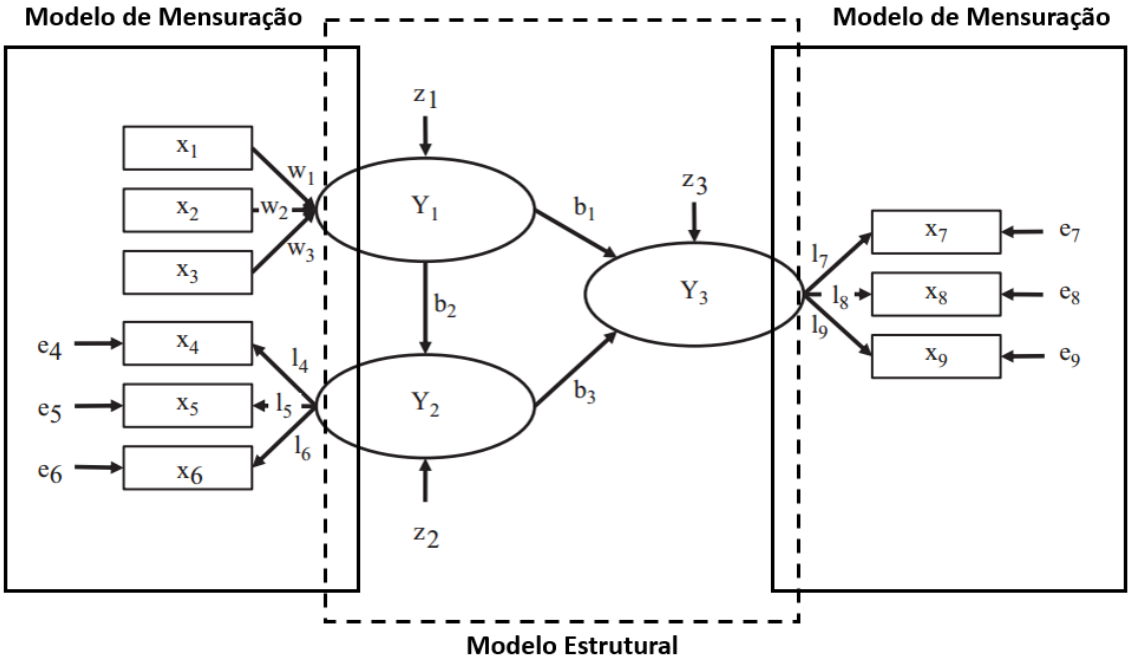

Figura 1 – Modelo SEM com variáveis latentes, indicadores e seus relacionamentos. No destaque, os modelos estrutural e de mensuração. Adaptado de Sarstedt, Ringle e Hair (2017).

res, e dos relacionamentos entre construtos. Os construtos, ou variáveis latentes, podem ser entendidos como elementos que representam as variáveis conceituais em um modelo teórico proposto por um pesquisador. Por representarem conceitos cuja observação não é direta, os construtos são medidos indiretamente por meio de indicadores, ou variáveis manifestas, ou ainda itens, sendo esses medidos diretamente. No modelo SEM, os construtos são representados por círculos (Y_1 a Y_3) e os indicadores (x_1 a x_9), por retângulos, sendo conectados por setas. O sentido das setas informa a natureza da contribuição dos indicadores para a formação do construto, o que será detalhado mais à frente.

Um modelo SEM consiste de dois elementos: modelo estrutural e modelo de validação, ou mensuração (Figura 1).

3.1.1.1 Modelo Estrutural

Também chamado de modelo interno, o modelo estrutural se expressa na disposição dos construtos e no relacionamento entre eles, representando as hipóteses e seus relacionamentos com a teoria sendo testada, tomando como base a literatura, a lógica e as experiências práticas do pesquisador.

No modelo as variáveis independentes, ou variáveis latentes exógenas, ou ainda preditores, são dispostas à esquerda (Y_1, na Figura 1) e as variáveis dependentes, ou variáveis latentes endógenas, são dispostas à direita (Y_3, na Figura 1). As variáveis latentes endógenas recebem setas que partem das variáveis exógenas. Variáveis latentes que operam como independentes e dependentes são representadas no meio do diagrama (Y_2, na Figura

1) (HAIR et al., 2016; SARSTEDT; RINGLE; HAIR, 2017).

Uma vez estabelecida a sequência dos construtos, ou variáveis latentes, os relacionamentos entre eles são então representados como setas, apontando para a direita, indicando que os construtos à esquerda predizem os construtos à direita. Os relacionamentos podem ser chamados de causais, quando houver teoria estrutural que fundamente essa relação causal.

Ainda no modelo interno, a força da relação entre os construtos é indicada por coeficientes (b_1 a b_3), calculados pela regressão de cada variável latente endógena em seu construto predecessor direto. Os indicadores z_1 a z_3 representam a variância não capturada pelos construtos antecedentes.

Num modelo estrutural o relacionamento direto entre duas variáveis latentes, indicado por uma seta simples, é chamado de efeito direto. Um relacionamento que envolva uma sequência de relacionamentos com pelo mesmo um construto interveniente é chamado de efeito indireto. A intervenção de um terceiro construto no relacionamento de outros dois construtos produz um efeito mediador, explicando ou tornando mais compreensível o relacionamento direto entre duas variáveis latentes (HAIR et al., 2016).

A disposição das variáveis latentes é discutida com profundidade por Hair et al. (2016).

3.1.1.2 Modelo de Mensuração

O relacionamento entre a variável latente e os seus indicadores é representado no modelo de validação, ou modelo externo ou, ainda, modelo de mensuração. A maneira como os construtos são medidos deve ser bem fundamentada na teoria, de forma que os testes de hipóteses envolvendo os relacionamentos estruturais entre as variáveis latentes sejam tão confiáveis e válidos quanto os indicadores.

O sentido das setas entre o construto e seus indicadores diferencia os modelos de validação entre reflexivos e formativos. Um modelo reflexivo representa efeitos ou manifestações de um determinado construto, o que é indicado por setas cujo sentido parte do construto para os indicadores. Nesse tipo de modelo, os indicadores de um construto podem ser entendidos como uma amostra significativa de todos os itens disponíveis no domínio conceitual do construto, possuindo alta correlação entre si. O relacionamento entre variáveis latentes (Y_2 e Y_3, na Figura 1) e indicadores reflexivos (x_4 a x_9, na Figura 1) pode ser formalmente expresso como na equação 3.1:

$$x = lY + e, \tag{3.1}$$

onde x é o coeficiente do indicador, Y é o coeficiente da variável latente, l é a carga

externa, que mede a força do relacionamento entre x e Y, e e mede o erro randômico (SARSTEDT; RINGLE; HAIR, 2017). A carga externa (ou *outer loading*), é calculada como o coeficiente de regressão da Equação 3.1.

Já no modelo formativo, a direção das setas parte dos indicadores para o construto. Nesse modelo, os indicadores se combinam linearmente para formar o construto e sua contribuição para a composição do construto deve ser diferenciada como causal ou como composta. No caso de construtos medidos por itens causais uma medida de erro deve ser acrescentada, indicando que causas não contempladas podem contribuir para a formação do construto. No modelo representado pela Figura 1, se os itens x_1 a x_3 forem modelados como indicadores causais, o erro z_1 deve ser levado em consideração. O modelo de mensuração de indicadores causais pode ser formalmente descrito como na equação 3.2:

$$Y = \sum_{k=1}^{K} w_k \cdot x_k + z, \tag{3.2}$$

onde w_k indica a contribuição dos indicadores $x_k (k = 1, \cdots, K)$ para o construto Y e z representa o erro associado ao construto Y (SARSTEDT; RINGLE; HAIR, 2017). No modelo formativo, o peso do indicador é chamado de peso externo (do inglês *outer weight*).

Uma segunda forma de combinar indicadores formativos é assumir que esses indicadores definem completamente o construto, considerando como zero o valor do erro z. Esse modelo de mensuração é chamado de modelo de indicadores compostos e pode ser formalmente descrito como na equação 3.3:

$$Y = \sum_{k=1}^{K} w_k \cdot x_k, \tag{3.3}$$

onde w_k indica o peso da contribuição dos indicadores $x_k (k = 1, \cdots, K)$ para o construto Y (SARSTEDT; RINGLE; HAIR, 2017).

3.1.2 Cálculo de pesos, cargas e coeficientes pelo PLS-SEM

O algoritmo PLS-SEM calcula e utiliza os coeficientes das variáveis latentes (Y_1 a Y_3, na Figura 1) como substitutos dos valores dos indicadores. Os coeficientes das variáveis latentes exógenas (preditoras) são estimados como combinações lineares exatas dos valores dos seus indicadores, de modo que a combinação resultante capture a maioria da variância desses indicadores e ajude a prever os indicadores das variáveis endógenas (SARSTEDT; RINGLE; HAIR, 2017). Como exemplo, na Figura 1 o valor da variável latente Y_1 é calculado como uma combinação linear dos indicadores x_1, x_2 e x_3.

O algoritmo PLS-SEM (Algoritmo 1) pode ser representado da seguinte maneira:

Algorithm 1: Algoritmo PLS-SEM. Adaptado de Henseler, Ringle e Sarstedt (2012) e Latan e Noonan (2017).

1 **Inicialização**
2 **Estágio 1:** cálculo iterativo de pesos (b_1 a b_3, na Figura 1) e de coeficientes (Y_1 a Y_3, na Figura 1) de variáveis latentes
3 **while** *não convergir* **do**
4 **Passo 1:** Pesos internos
5 $$b_{ji} = \begin{cases} cov(Y_j; Y_i) & \text{se } Y_j \text{ e } Y_i \text{ forem adjacentes} \\ 0 & \text{outros casos} \end{cases}$$
6 **Passo 2:** Aproximação interna
7 $\tilde{Y}_j = \sum_i b_{ji} Y_i$
8 **Passo 3:** Pesos externos
9 $\tilde{Y}_{jn} = \sum_{kj} \tilde{w}_{k_j} x_{k_j n} + d_{jn}$ (Modo A)
10 $x_{kjn} = \tilde{w}_{k_j} \tilde{Y}_{jn} + e_{k_j n}$ (Modo B)
11 **Passo 4:** Aproximação externa
12 $Y_{jn} = \sum_{kj} \tilde{w}_{k_j} x_{k_j n}$
13 **Estágio 2:** cálculo dos pesos externos, cargas externas e coeficientes de variáveis latentes

Como apresentado por Latan e Noonan (2017), o algoritmo PLS-SEM é **inicializado** (linha 1 do Algoritmo 1) pela definição preliminar dos coeficientes das variáveis latentes pela atribuição do peso 1 (um) para todos os indicadores do modelo de mensuração. Na prática, para essa definição preliminar de coeficientes das variáveis latentes, o algoritmo executa o Passo 4 do Estágio 1 (linha 12 do Algoritmo 1), calculando os coeficientes das variáveis latentes pelo somatório do produto do coeficiente normalizado ($x_{k_j n}$) pelo peso ($\tilde{w}_{k_j}$) (arbitrariamente definido como 1) de seus indicadores.

O **Estágio 1** estima os pesos internos (b_1 a b_3) e os coeficientes das variáveis latentes (Y_1 a Y_3) por um procedimento iterativo de quatro passos (linhas 4 a 12 do Algoritmo 1). No **Passo 1** do **Estágio 1** (linha 4 do Algoritmo 1) calculam-se os pesos internos (b_{ji}), obtidos pelo resultado da covariância entre os coeficientes normalizados das variáveis latentes dependentes (Y_j), ou endógenas, e os coeficientes normalizados das variáveis latentes independentes (Y_i), ou exógenas. O peso interno será 0 (zero) para variáveis latentes não conectadas.

No **Passo 2** do **Estágio 1** (linha 6 do Algoritmo 1) os coeficientes das variáveis latentes são atualizados com base nos novos pesos internos obtidos no **Passo 1**. Para as variáveis latentes exógenas Y_i (Y_1 e Y_2, na Figura 1) o novo coeficiente será obtido pelo produto do coeficiente da variável endógena Y_j (Y_3, na Figura 1) pelo peso da relação com a variável exógena Y_i (b_1 ou b_3, na Figura 1). Como exemplo, na Figura 1 os novos

valores de Y_1 e Y_2 serão obtidos por $Y_1 = Y_3 \cdot b_1$ e $Y_2 = Y_3 \cdot b_3$. Ainda no **Passo 2**, os novos coeficientes das variáveis latentes endógenas Y_j (Y_3, na Figura 1) serão calculados como o somatório do produto entre os coeficientes das variáveis latentes exógenas Y_i (Y_1 ou Y_2, na Figura 1) e os pesos das relações com a variável endógena Y_j (b_1 ou b_3, na Figura 1). Como exemplo, na Figura 1 o novo coeficiente da variável Y_3 é calculado como $\tilde{Y_3} = Y_1 \cdot b_1 + Y_2 \cdot b_3$. Os novos coeficientes das variáveis latentes exógenas e endógenas são, então, normalizados.

No **Passo 3** do **Estágio 1** (linha 8 do Algoritmo 1), novos pesos (w_1 a w_3 e l_4 a l_9, na Figura 1) são calculados para os indicadores dos modelos de mensuração (x_1 a x_9, na Figura 1) indicando a força do relacionamento dos indicadores com as variáveis latentes. Para o cálculo o algoritmo PLS-SEM usa dois modos de estimação: o **Modo A**, usado por padrão em relações reflexivas (por exemplo, Y_2 em relação a x_4, x_5 e x_6 na Figura 1), e o **Modo B**, usado por padrão em relações formativas (por exemplo, Y_1 em relação a x_1, x_2 e x_3 na Figura 1). No **Modo A** (formações reflexivas) as cargas (l_4 a l_9, na Figura 1) são obtidas como a correlação bivariada entre cada indicador e a variável latente. No **Modo B** (formações formativas) os pesos (w_1 a w_3, na Figura 1) são obtidos pela regressão de cada variável latente em seus indicadores. Nas linhas 9 e 10 do Algoritmo 1 x_{k_jn} representa os dados dos indicadores $k(k = 1, \cdots, K)$ das variáveis latentes $j(j = 1, \cdots, J)$ e observações $n(n = 1, \cdots, N)$, $\tilde{Y}_{jn}$ representa os coeficientes das variáveis latentes obtidos no **Passo 2**, $\tilde{w}_{k_j}$ são os pesos externos obtidos no **Passo 3**, d_{jn} representa o termo de erro da regressão bivariada e e_{k_jn} representa o termo de erro da regressão múltipla.

O **Passo 4** do **Estágio 1** (linha 11 do Algoritmo 1) combina linearmente os pesos $\tilde{w}_{k_j}$ e os coeficientes x_{k_jn} dos indicadores, obtidos no **Passo 3**, para calcular os coeficientes das variáveis latentes Y_{jn}, normalizando os valores ao final do cálculo. O **Estágio 1** termina quando os pesos obtidos no **Passo 3** apresentam baixa variação de uma iteração para a próxima (variação de 1×10^{-7}) ou quando o número máximo de iterações for atingido (por padrão, 300).

No **Estágio 2** os pesos das variáveis latentes (b_1 a b_3, na Figura 1) são calculados pelo método dos mínimos quadrados (OLS) baseado em regressão tomando como base os coeficientes das variáveis latentes calculados no **Estágio 1** (HENSELER; RINGLE; SARSTEDT, 2012). Os pesos das variáveis latentes (b_1 a b_3, na Figura 1) corresponderão aos coeficientes de regressão linear. O coeficiente de determinação (R^2) também é retornado nesse passo, e corresponderá ao valor final do coeficiente da variável latente endógena (Y_3, na Figura 1).

3.1.3 Validação do Modelo

Como o método PLS-SEM visa à predição de um conjunto de relacionamentos hipotéticos para maximizar a variância explicada das variáveis dependentes, a adequação do modelo é medida pela discrepância entre os valores dos indicadores, ou dos valores aproximados das variáveis latentes, e os valores preditos pelo modelo. Assim, o modo de avaliar um modelo é por meio de medidas que avaliem seu potencial preditivo. As medidas de avaliação do modelo separam-se entre aquelas que avaliam o modelo estrutural e as que avaliam o modelo de mensuração (HAIR et al., 2016).

Sarstedt, Ringle e Hair (2017) seguem Hair et al. (2016) na proposta de um procedimento de validação dos resultados do PLS-SEM consistindo de três estágios: começando pela validação do modelo de mensuração reflexivo, passando pela validação do modelo de mensuração formativo, e, caso haja suporte para a qualidade da mensuração, chegando à validação do modelo estrutural. O procedimento é descrito nas próximas três subseções, que tomam como base os trabalhos de Hair et al. (2016) e Sarstedt, Ringle e Hair (2017).

3.1.3.1 Medidas de Avaliação do Modelo de Mensuração Reflexivo

A avaliação do modelo de mensuração reflexivo começa pelo exame das cargas dos indicadores (l_4 a l_9, na Figura 1). Uma vez que as cargas de indicadores reflexivos são obtidas como a correlação bivariada entre cada indicador e a variável latente (linha 9 do Algoritmo 1), valores acima de 0,70 indicam que o a variável latente explica mais de 50% da variância do indicador, o que é considerado um grau satisfatório de confiabilidade.

A confiabilidade interna do construto reflexivo é o passo seguinte na avaliação do modelo de mensuração, podendo ser medida no limite inferior pelo índice Alfa de Cronbach α (Equação 3.4) e no limite superior, pelo índice de confiabilidade composta ρ_c (Equação 3.5):

$$\alpha = \frac{K \cdot \bar{r}}{[1 + (K-1) \cdot \bar{r}]}, \tag{3.4}$$

onde K indica o número de indicadores da variável latente e $\bar{r}$ representa a média não redundante do coeficiente de correlação do indicador, e

$$\rho_c = \frac{(\sum_{k=1}^{K} l_k)^2}{(\sum_{k=1}^{K} l_k)^2 + \sum_{k=1}^{K} var(e_k)}, \tag{3.5}$$

onde l_k indica o valor padronizado da carga do indicador k de um construto com K indicadores, e_k representa o erro de mensuração do indicador k e $var(e_k)$ representa a variância do erro de medida, calculada como $var(e_k) = 1 - l_k^2$. Valores aceitáveis a bons para α e ρ_c são 0,60 a 0,95, indicando graus crescentes de confiabilidade.

O passo seguinte visa à medida da validade convergente, que é a extensão em que um indicador se correlaciona positivamente com indicadores alternativos para o mesmo construto. Considerando que os indicadores de um construto reflexivo são tratados como diferentes abordagens para medir o mesmo construto, os indicadores desse construto devem convergir ou compartilhar uma alta proporção de variância. Uma medida comum de validade convergente é a Variância Média Extraída (AVE), equivalente à comunalidade do construto. Valores de 0,5 ou maiores indicam que o construto explica mais da metade da variância de seus indicadores (ZWICKER; SOUZA; BIDO, 2008; HAIR et al., 2016). A AVE pode ser obtida por (Equação 3.6):

$$\text{AVE} = \frac{(\sum_{k=1}^{K} l_k)^2}{K}, \tag{3.6}$$

onde l_k indica o valor padronizado da carga do indicador k de um construto com K indicadores.

No passo final da avaliação do modelo de mensuração reflexivo, mede-se a validade discriminante, que pode ser entendida como a extensão na qual uma variável latente é de fato distinta de outras variáveis latentes por padrões empíricos. Isso significa que se a validade discriminante de um construto for estabelecida, esse construto é único e captura um fenômeno não representado por outros construtos do modelo.

A validade discriminante pode ser medida pelo exame das cargas cruzadas dos indicadores, no sentido de que a carga de um item num construto deve ser maior que sua carga em outros construtos. Uma segunda forma de medir a validade discriminante é o critério de Fornell-Larcker, que estabelece que a raiz quadrada da Variância Média Extraída (AVE) de cada variável latente deve ser maior que maior correlação entre variáveis latentes (ZWICKER; SOUZA; BIDO, 2008; HAIR et al., 2016).

Um método mais recente para a avaliação da validade discriminante é a taxa de correlação heterotraço-monotraço (HTMT - *heterotrait-monotrait ratio*), que pode ser compreendida como o valor médio das correlações de um indicador entre variáveis latentes relativo à média geométrica das correlações de indicadores medindo a mesma variável latente. O cálculo da taxa HTMT das variáveis latentes Y_i e Y_j é obtido por (Equação 3.7):

$$HTMT_{ij} = \frac{\frac{1}{K_i K_j} \sum_{g=1}^{K_i} \sum_{h=1}^{K_j} r_{ig,jh}}{(\frac{2}{K_i(K_i-1)} \cdot \sum_{g=1}^{K_i-1} \sum_{h=g+1}^{K_i} r_{ig,ih} \cdot \frac{2}{K_j(K_j-1)} \cdot \sum_{g=1}^{K_j-1} \sum_{h=g+1}^{K_j} r_{jg,jh})^{\frac{1}{2}}} \tag{3.7}$$

onde K_i e K_j representam os indicadores das variáveis latentes Y_i e Y_j, $r_{ig,jh}$ representa as correlações de indicadores dentre e através dos modelos de mensuração dos construtos

Y_i e Y_j. Valores de HTMT acima de 0,85 em um modelo onde as variáveis latentes são conceitualmente diferentes indicam problemas de validade discriminante.

3.1.3.2 Medidas de Avaliação do Modelo de Mensuração Formativo

Porque pesos de indicadores de variáveis latentes formativas são calculados pela regressão de cada variável latente em seus indicadores (linha 10 do Algoritmo 1), a avaliação do modelo de mensuração formativo requer instrumentos diferentes dos utilizados na avaliação do modelo de mensuração reflexivo.

O primeiro instrumento é a verificação da validade convergente, já detalhada anteriormente, mas agora compreendida da seguinte maneira: a validade convergente de construtos formativos é determinada como a extensão pela qual esse construto se correlaciona com outro construto reflexivo de um único item que capture o mesmo conceito. Esse construto reflexivo de um único item é utilizado nas Ciências Sociais como um recurso de análise redundante.

Outro instrumento é a validação da colinearidade, pela obtenção do índice VIF em cada um dos indicadores das variáveis latentes formativas. O calculo do índice VIF (Equação 3.8) consiste na aplicação de uma regressão múltipla de cada indicador em todos os outros indicadores da mesma variável latente:

$$VIF_k = \frac{1}{1 - R_k^2}, \tag{3.8}$$

onde R_k^2 corresponde ao coeficiente de determinação da k-ésima regressão, para o cálculo do índice VIF do k-ésimo indicador. Valores altos de R^2 indicam que a variância do indicador pode ser explicada pelos outros indicadores do mesmo construto, configurando a colinearidade do indicador. Valores acima de 5 indicam colinearidade de indicadores (HAIR et al., 2016).

O passo final da validação do modelo de mensuração formativo é a avaliação da contribuição relativa de cada indicador na formação do construto. Valores normalizados dos pesos dos indicadores próximos a 0 indicam relacionamentos fracos, e valores próximos a 1 ou a -1 representam relacionamentos fortes. Ainda na validação final, a significância estatística do indicador deve ser avaliada. O procedimento de *bootstrapping* (HAIR et al., 2016; STREUKENS; LEROI-WERELDS, 2016) permite a computação dos erros padrão e a significância do peso do indicador. Se o peso do indicador for estatisticamente significante, o indicador é mantido. Se o peso do indicador não for estatisticamente significante mas sua carga for igual ou maior que 0,50, o indicador poderá ser mantido se houver suporte teórico para isso. O indicador deve ser retirado no caso de peso inferior a 0,50 e não haver significância estatística do peso (HAIR et al., 2016). Henseler, Ringle e Sarstedt (2012) afirmam que, no caso de uso de vários indicadores formativos e na presença de

alguns pesos não significativos estatisticamente, os indicadores devem ser agrupados em outros construtos, caso haja suporte teórico.

3.1.3.3 Medidas de Avaliação do Modelo Estrutural

Após a conclusão da avaliação dos modelos de mensuração, o modelo estrutural é avaliado quanto a problemas de colinearidade entre as variáveis latentes e quanto à sua capacidade preditiva. A avaliação de problemas de colinearidade dos construtos usa o mesmo instrumento utilizado na avaliação da colinearidade das variáveis latentes formativas, o índice VIF (Equação 3.8), com a diferença que ao invés de indicadores, o *input* da equação serão as variáveis latentes exógenas. Valores de VIF superiores a 5 indicarão colinearidade entre as variáveis latentes preditoras (SARSTEDT; RINGLE; HAIR, 2017).

Quanto à avaliação da capacidade preditiva do modelo, três critérios devem ser considerados: o coeficiente de determinação (R^2), a redundância de validação cruzada(Q^2) e os coeficientes do modelo. O coeficiente de determinação R^2 indica a variância explicada nas variáveis latentes endógenas, significando a acurácia do modelo preditivo. Em geral, valores de 0,75 , 0,50 e 0,25 são considerados, respectivamente, substancial, moderado e fraco, devendo, no entanto, ser considerado o contexto da análise. Sarstedt, Ringle e Hair (2017) explicam ainda que a omissão de uma variável latente exógena pode ser um instrumento para a avaliação de seu impacto no modelo preditivo, que pode ser medido por (Equação 3.9):

$$f^2 = \frac{R^2_{incluido} - R^2_{excluido}}{1 - R^2_{incluido}}, \tag{3.9}$$

onde $R^2_{incluido}$ e $R^2_{excluido}$ correspondem aos valores de R^2 calculados com e sem a variável latente analisada. Valores de f^2 inferiores a 0,02 indicam que não há efeito da variável latente no modelo, e valores de 0,02 , 0,15 e 0,35 são considerados, respectivamente, efeitos pequeno, médio e grande da variável latente no modelo.

Outra validação do modelo estrutural é chamada de Q^2 de Stone-Geisser, que mede a relevância preditiva do modelo por um processo iterativo de omissão “às cegas” de pontos (dados) dos indicadores de variáveis latentes reflexivas ou de variáveis latentes endógenas com somente um indicador para que se verifique se o algoritmo PLS-SEM prediz acuradamente os pontos omissos usando os pontos restantes. Os pontos omissos são tratados pelo algoritmo PLS-SEM como valores ausentes, sendo repostos pelo critério da média (HAIR et al., 2016). Se o valor predito for próximo ao real, com um erro de predição baixo, o modelo apresentará alta acurácia preditiva.

Por fim, mas não menos importante é a validação dos pesos dos relacionamentos entre as variáveis latentes, cujos valores padronizados variam entre -1 e +1. Valores próximos de +1 ou -1 indicam relacionamentos fortes, que em geral são estatisticamente

significantes. Quanto mais próximos os pesos forem de 0, mais fracos serão os relacionamentos. No entanto, o procedimento de *bootstraping* permitirá a obtenção do erro padrão e o cálculo do valor t empírico. Para estimar a significância do relacionamento entre as variáveis latentes Y_1 e Y_3, da Figura 1, o cálculo a ser feito é (Equação 3.10):

$$t = \frac{b_1}{se^*_{b_1}} \tag{3.10}$$

onde b_1 corresponde ao peso do relacionamento entre as variáveis latentes Y_1 e Y_3, da Figura 1, e $se^*_{b_1}$ corresponde ao erro padrão do peso b_1 obtido pelo método de *bootstrapping*. Segundo Hair et al. (2016), os quantis de uma distribuição normal podem ser usados como valores críticos contra os quais o valor t será comparado. Valores comumente usados são 1,65 (nível de significância de 10%), 1,96 (nível de significância de 5%) e 2,57 (nível de significância de 1%). Em estudos exploratórios pode-se assumir um nível de significância de 10%. O valor p é também utilizado em conjunto com o valor t como indicador da significância do relacionamento entre duas variáveis latentes.

3.2 Dimensões geométricas de Complexidade dos Dados

As tentativas de se estabelecer medidas de complexidade para sistemas e seus dados têm mostrado que medidas indiretas, tais como complexidade algorítmica de Kolgomorov (1965) e a entropia de Shannon (1948), são a abordagem viável, embora a primeira não seja computável (veja Bialek, Nemenman e Tishby (2001) e Boschetti (2008)). Assim, encontram espaço as abordagens que analisam a complexidade de dados por suas dimensões geométrica, estatística e de qualidade (HO; BASU; LAW, 2006). Nesta seção serão exploradas dimensões dos dados para as quais a tarefa de classificação de dados é sensível.

A tarefa de classificação de dados tem recebido destaque em publicações em que a Complexidade de Dados tangencia a Mineração de Dados, possivelmente devido à sensibilidade dessa tarefa a características geométricas dos dados. Ho (HO; BASU, 2002; HO; BASU; LAW, 2006) enumera três dificuldades para as tarefas de classificação: a) ambiguidade de classes, que ocorre quando os atributos do conjunto de dados não são suficientes para que um algoritmo de classificação, qualquer que seja, faça a distinção entre as classes, seja porque as classes não estão claramente definidas ou porque os atributos não são informativos o suficiente para a separação das classes; b) complexidade de fronteira, cujo grau pode ser medido pela quantidade de informação necessária para se descrever o limite entre as classes em um conjunto de dados; e c) esparsamento da amostra e dimensionalidade do espaço de atributos, que ocorre quando a capacidade de generalização de um classificador é prejudicada por amostras que representem de maneira insuficiente o conjunto de dados, ainda mais quando o espaço de atributos é grande, aumentando a variabilidade da região

de decisão do classificador. As medidas de complexidade geométrica utilizadas por Ho capturam a distribuição espacial e a estrutura dos dados.

A preocupação com as sensibilidades dos algoritmos de classificação relacionadas a qualidade ou complexidade é compartilhada na literatura por outros pesquisadores, tais como Sánchez, Mollineda e Sotoca (2007), que discutem o efeito negativo que algumas dimensões da complexidade dos dados exercem sobre o algoritmo de aprendizagem supervisionada k-NN, incluindo a alta dimensionalidade dos dados, a sobreposição e a densidade de classes. Garcia, Carvalho e Lorena (2015) exploram o efeito da presença de ruídos no componente de classe de um conjunto de dados, usando medidas de sobreposição e separabilidade de classes, geometria e topologia, e representação estrutural dos dados. Zubek e Plewczynski (2016) apresentam uma medida de complexidade gráfica com aplicação direta na redução dos dados da etapa de treinamento de classificadores. Barella et al. (2018) aplicam medidas de complexidade em conjuntos de dados reais e artificiais para identificar o efeito da sobreposição de classes em tarefas de classificação onde as classes estão desbalanceadas.

As dimensões de interesse da presente pesquisa são as apresentadas por Lorena et al. (2019), baseadas originalmente nas pesquisas de Ho e Basu (2002) e Lorena e Souto (2015), e agrupadas em categorias, conforme apresentado na Tabela 1.

Faz-se necessário destacar que as dimensões da complexidade dos dados, conforme definidas originalmente, nem sempre mantêm uma relação direta com a complexidade do conjunto de dados. Por exemplo, quanto maior o valor do indicador F2 (volume da região de sobreposição) maior a complexidade do conjunto de dados, o que pode ser entendido como uma correlação positiva (+) com a complexidade do conjunto de dados, mas quanto maior o valor do indicador F3 (eficiência do atributo) menor a complexidade do conjunto de dados analisado, o que pode ser representado por uma correlação negativa (-). A coluna "Correlação com Complexidade" indica se a correlação do indicador com a complexidade do conjunto de dados é positiva (+) ou negativa (-). As dimensões de complexidade geométrica dos dados são detalhadas a seguir, com base o trabalho de Lorena et al. (2019).

3.2.1 Medidas de atributos

Nessa categoria avalia-se o poder discriminativo de atributos, tratando-se como menos complexos os conjuntos de dados que possuírem ao menos um atributo discriminativo (HO; BASU, 2000).

Taxa máxima discriminante de Fisher ($F1$)

F1 mede a sobreposição entre os valores dos atributos de classes diferentes, podendo ser aplicada a classificações binárias ou multi-classes. O cálculo de F1 apresentado por

Tabela 1 – Dimensões geométricas de complexidade dos dados. Fonte: Lorena et al. (2019)

Dimensão	Nome	Correlação com Complexidade
	Medidas de atributos	
F1	Taxa máxima discriminante de Fisher	(+)
F1v	Taxa máxima discriminante vetor-direcional de Fisher	(-)
F2	Volume da região de sobreposição	(+)
F3	Eficiência máxima individual do atributo	(-)
F4	Eficiência coletiva do atributo	(-)
	Medidas de vizinhança	
N1	Fração de pontos na fronteira da classe	(+)
N2	Taxa média de distância NN extra/intra classe	(+)
N3	Taxa de erro *leave-one-out* do classificador 1NN	(+)
N4	Não-linearidade do classificador 1-NN	(+)
T1	Fração de hiperesferas cobrindo os dados	(+)
LSC	Cardinalidade média do conjunto local	(+)
	Medidas de linearidade	
L1	Soma do erro de distância por programação linear	(+)
L2	Taxa de erro de classificador linear	(+)
L3	Não-linearidade de um classificador linear	(+)
	Medidas de dimensionalidade	
T2	Número médio de pontos por dimensão	(-)
T3	Número médio de pontos por dimensões PCA	(-)
T4	Taxa de dimensões PCA em relação às originais	(+)
	Medidas de desbalanceamento de classes	
C1	Entropia das proporções de classe	(-)
C2	Taxa de desbalanceamento	(+)
	Medidas de rede	
Density	Densidade média da rede	(-)
ClsCoef	Coeficiente de agrupamento	(-)
Hubs	Indice de pontos centrais	(-)

Lorena et al. (2019) já traz resultados normalizados, indicando que valores próximos a 1 representam conjuntos de dados com poucos atributos discriminantes e, por isso, mais complexos.

F1 pode ser calculada como na Equação 3.11:

$$F1 = \frac{1}{1 + max_{i=1}^{m} r_{f_i}}, \tag{3.11}$$

onde r_{f_i} é a taxa discriminante para cada atributo f_i, e é calculada por Lorena et al. (2019) como na Equação 3.12:

$$r_{f_i} = \frac{\sum_{j=1}^{n_c} n_{c_j}(\mu_{c_j}^{f_i} - \mu^{f_i})^2}{\sum_{j=1}^{n_c}\sum_{l=1}^{n_{c_j}}(x_{li}^{j} - \mu_{c_j}^{f_i})^2}, \tag{3.12}$$

onde n_{c_j} corresponde ao número de objetos na classe c_j, $\mu_{c_j}^{f_i}$ corresponde à média do atributo f_i para a classe c_j, μ^{f_i} corresponde à média dos valores f_i para todas as classes, e x_{li}^{j} corresponde ao valor individual do atributo f_i para um objeto da classe c_j.

Taxa máxima discriminante vetor-direcional de Fisher ($F1v$)

$F1v$ é complementar a $F1$, buscando um vetor que separe as duas classes após os objetos serem projetados nele. Quanto maior o valor de $F1v$, mais simples é o conjunto de dados (LORENA et al., 2019, p.3). A medida, já normalizada, é calculada pela Equação 3.13:

$$F1v = \frac{1}{1 + dF}, \tag{3.13}$$

onde dF corresponde ao critério direcional de Fisher, definido pela Equação 3.14:

$$dF = \frac{d^t B d}{d^t W d}, \tag{3.14}$$

onde d corresponde ao vetor direcional no qual cada dado é projetado para maximizar a separação das classes, B corresponde à matriz de dispersão inter-classes, W corresponde à matriz de dispersão intra-classes, e são calculados respectivamente pelas Equações 3.15, 3.16 e 3.17:

$$d = W^{-1}(\mu_{c_1} - \mu_{c_2}), \tag{3.15}$$

onde μ_{c_i} corresponde ao centroide da classe c_i e W^{-1} corresponde ao pseudo inverso de W.

$$B = (\mu_{c_1} - \mu_{c_2})(\mu_{c_1} - \mu_{c_2})^t, \tag{3.16}$$

$$W = p_{c_1}\Sigma_{c_1} + p_{c_2}\Sigma_{c_2} \tag{3.17}$$

onde p_{c_i} corresponde à proporção de objetos na classe c_i e Σ_{ci} corresponde à matriz de dispersão da classe c_i.

Volume da região de sobreposição ($F2$)

$F2$ calcula a sobreposição das distribuições dos valores dos atributos dentro das classes. Para cada atributo f_i obtêm-se os valores mínimo e máximo nas classes, e a região

de sobreposição é então calculada e normalizada pelo intervalo de valores de ambas as classes. Os valores são então multiplicados, como mostra a Equação 3.18:

$$F2 = \prod_{i}^{m} \frac{overlap(f_i)}{range(f_i)} = \prod_{i}^{m} \frac{max\{0, min\ max(f_i) - max\ min(f_i)\}}{max\ max(f_i) - min\ min(f_i)}, \tag{3.18}$$

onde:

$min\ max(f_i) = min(max(f_i^{c_1}), max(f_i^{c_2}))$,

$max\ min(f_i) = max(min(f_i^{c_1}), min(f_i^{c_2}))$,

$max\ max(f_i) = max(max(f_i^{c_1}), max(f_i^{c_2}))$,

$min\ min(f_i) = min(min(f_i^{c_1}), min(f_i^{c_2}))$. Os valores $max(f_i^{c_j})$ e $min(f_i^{c_j})$ correspondem aos valores máximo e mínimo de cada atributo numa classe $c_j \in \{1, 2\}$. O numerador torna-se zero quando os intervalos de valor por classe são disjuntos para pelo menos um atributo.

Eficiência máxima individual do atributo ($F3$)

$F3$ estima a eficiência de cada atributo em separar as classes, levando em consideração o maior valor encontrado entre m atributos. Para cada atributo, verifica-se a existência de sobreposição entre as classes, considerando-as ambíguas na região em que houver sobreposição. Um problema é considerado simples se houver ao menos um atributo que apresente baixa ambiguidade entre classes. $F3$ pode ser calculado como na Equação 3.19:

$$F3 = \max_{i=1}^{m} \frac{n - n_0(f_i)}{n}, \tag{3.19}$$

onde $n_0(f_i)$ corresponde ao número de objetos que estão na região de sobreposição para o atributo f_i e pode ser calculado como na Equação 3.20:

$$n_0(f_i) = \sum_{j=1}^{n} I(x_{ji} > max\ min(f_i) \wedge x_{ji} < min\ max(f_i)), \tag{3.20}$$

onde I retorna 1 se seus argumentos forem verdadeiros e retorna 0 nos demais casos. $max\ min(f_i)$ e $min\ max(f_i)$ são definidos da mesma forma que em $F2$. Nessa medida, quanto maior o valor, mais simples o problema (LORENA et al., 2019, p.6).

Eficiência coletiva do atributo ($F4$)

$F4$ aplica sucessivamente um procedimento similar ao adotado em $F3$. Inicialmente identifica-se o atributo mais discriminativo, eliminando-se em seguida todos os objetos que possam ser separados por esse atributo. O procedimento é repetido até que todos os atributos tenham sido considerados e não restem objetos que possam ser eliminados. O

resultado de $F4$ é calculado pela taxa de objetos que não foram discriminados em relação ao total de objetos, como na Equação 3.21:

$$F4 = \frac{n_o(f_{min}(T_l))}{n}, \tag{3.21}$$

onde l corresponde ao índice no intervalo $[1, m]$ do atributo capaz de discriminar todos os objetos em T, e $n_o(f_{min}(T_l))$ corresponde ao número de pontos na região de sobreposição do atributo f_{min} para o *dataset* na l-ésima rodada (T_l). Para a i-ésima iteração de $F4$ o atributo mais discriminativo no *dataset* pode ser calculado como na Equação 3.22:

$$f_{min}(T_i) = \left\{ f_j | \min_{j=1}^{m}(n_o(f_j)) \right\}_{T_i}, \tag{3.22}$$

onde $n_o(f_j$ é calculado como na Equação 3.20, e o *dataset* a cada rodada pode ser definido como nas Equações 3.23 e 3.24:

$$T1 = T, \tag{3.23}$$

$$T_i = T_{i-1} - \{x_j | x_{ji} < max\ min(f_{min}(T_{i-1})) \vee x_{ji} > min\ max(f_{min}(T_{i-1}))\}. \tag{3.24}$$

Valores maiores de F4 indicam um problema mais simples (LORENA et al., 2019, p.6).

3.2.2 Medidas de linearidade

Nessa categoria tenta-se quantificar a possibilidade de se separar as classes por um hiperplano baseado em *Support Vector Machine* (SVM), assumindo-se que um problema linearmente separável é mais simples que um problema que requeira um limite de decisão não linear (LORENA et al., 2019).

Soma do erro de distância por programação linear ($L1$)

$L1$ computa a separabilidade linear de classes pela computação da soma das distâncias dos objetos classificados incorretamente em relação ao hiperplano utilizado para sua separação. Em um problema linearmente separável, essa soma é zero. A Equação 3.25 expressa esse cálculo:

$$SumErrorDist = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i, \tag{3.25}$$

onde ε_i corresponde ao valor da distância de erro do objeto i erroneamente classificado. O valor de $L1$ pode ser calculado como na Equação 3.26:

$$L1 = 1 - \frac{1}{1 + SumErrorDist} = \frac{SumErrorDist}{1 + SumErrorDist}. \tag{3.26}$$

Taxa de erro de classificador linear ($L2$)

$L2$ computa a taxa de erros do classificador linear SVM, e pode ser obtido pela Equação 3.27:

$$L2 = \frac{\Sigma_{i=1}^{n} I(h(x_i) \neq y_i)}{n}, \tag{3.27}$$

onde $h(x)$ corresponde ao classificador linear obtido.

Não-linearidade de um classificador linear ($L3$)

$L3$ calcula a taxa de erro de um classificador linear testado em um conjunto de dados gerado a partir de objetos do conjunto de dados original. Cada objeto do conjunto de teste é obtido pela interpolação linear de dois objetos da mesma classe, escolhidos aleatoriamente no conjunto de dados original. O cálculo pode ser observado na Equação 3.28:

$$L3 = \frac{1}{l} \sum_{i=1}^{l} I(h_T(x'_i) \neq y'_i), \tag{3.28}$$

onde l corresponde ao número de objetos interpolados x'_i e sua respectiva classe y'_i, e $h_T(x)$ corresponde ao classificador linear induzido a partir dos objetos originais T. Valores altos para $L3$ indicam alta complexidade.

3.2.3 Medidas de vizinhança

As medidas dessa categoria tentam caracterizar a sobreposição de classes, capturar a forma da região de decisão e a estrutura interna das classes pela análise da vizinhança dos pontos. As distâncias entre pares de pontos são armazenadas em uma matriz, medidas pela distância de Gower.

Fração de pontos na fronteira da classe ($N1$)

$N1$ estima a complexidade e o tamanho de uma região de decisão necessária para separar os objetos de classes diferentes. Para tanto, uma Árvore Geradora Mínima (MST) é construída a partir dos dados originais, onde cada vértice corresponde a um objeto e as arestas são ponderadas de acordo com a distância entre os pontos. O valor de $N1$ representa o percentual de vértices incidentes para arestas conectando objetos de classes opostas na Árvore Geradora Mínima, como apresentado pela Equação 3.29:

$$N1 = \frac{1}{n} \sum_{i=1}^{n} I((x_i, x_j) \in MST \wedge y_i \neq y_j). \tag{3.29}$$

Taxa média de distância NN extra/intra classe ($N2$)

$N2$ calcula a taxa entre a soma das distâncias entre cada objeto e seu vizinho mais próximo da mesma classe e a soma de cada objeto e seu vizinho mais próximo de outra classe, como na Equação 3.30 e 3.31:

$$N2 = 1 - \frac{1}{1 + intra\_extra} = \frac{intra\_extra}{1 + intra\_extra}, \tag{3.30}$$

$$intra\_extra = \frac{\Sigma_{i=1}^{n} d(x_i, NN(x_i) \in y_i)}{\Sigma_{i=1}^{n} d(x_i, NN(x_i) \in y_j \neq y_i)}, \tag{3.31}$$

onde $d(x_i, NN(x_i) \in y_i)$ corresponde à distância do objeto x_i para seu vizinho mais próximo da mesma classe, e $d(x_i, NN(x_i) \in y_j \neq y_i)$ corresponde à distância do objeto x_i para seu vizinho mais próximo da outra classe. Valores menores de $N2$ indicam problemas mais simples.

Taxa de erro *leave-one-out* do classificador 1NN ($N3$)

$N3$ refere-se à taxa de erro do classificador k-Nearest Neighbors para k=1, estimado pelo procedimento *leave-one-out*, calculado como na Equação 3.32:

$$N3 = \frac{\Sigma_{i=1}^{n} I(NN(x_i) \neq y_i)}{n}, \tag{3.32}$$

onde $NN(x_i)$ corresponde à predição do classificador kNN para o objeto x_i.

Não-linearidade do classificador 1-NN ($N4$)

$N4$ é similar à medida $L3$, exceto por utilizar o classificador kNN ao invés de um preditor linear, e é calculado conforme apresentado na Equação 3.33:

$$N4 = \frac{1}{l} \sum_{i=1}^{l} I(NN_T(x'_i) \neq y'_i), \tag{3.33}$$

onde l corresponde ao número de objetos interpolados x'_i e sua respectiva classe y'_i, e $NN_T(x)$ corresponde ao classificador NN induzido a partir dos objetos originais T. Valores altos para $N4$ indicam alta complexidade.

Fração de hiperesferas cobrindo os dados ($T1$)

$T1$ computa a razão entre o número de hiperesferas e o número total de objetos do conjunto de dados. As hiperesferas são construídas como proposto no algoritmo de Lorena et al. (2019), e o valor de $T1$ é obtido como na Equação 3.34:

$$T1 = \frac{\sharp Hiperesferas(T)}{n}, \tag{3.34}$$

onde $\sharp Hiperesferas(T)$ corresponde ao número de hiperesferas necessário para cobrir om conjunto de dados. Quanto menor o número de hiperesferas cobrindo os dados, menor a complexidade desse conjunto de dados, indicando que os dados de uma mesma classe são densamente distribuídos e próximos uns aos outros.

Cardinalidade média do conjunto local (LSC)

Para o cálculo de LSC são considerados os conjuntos locais (LS), que são grupos de objetos de um conjunto de dados T cuja distância para um objeto x_i é menor que a distância do objeto x_i para seu inimigo mais próximo, como definido pela Equação 3.35:

$$LS(x_i) = \{x_j | d(x_i, x_j) < d(x_i, ne(x_i))\}, \tag{3.35}$$

onde $ne(x_i)$ corresponde ao inimigo mais próximo de x_i. A cardinalidade do LS de um objeto x_i indica sua proximidade para a região de decisão e a estreiteza do espaço entre as classes. A cardinalidade será menor para objetos separados de outras classes por uma margem estreita.

A cardinalidade média LSC pode ser calculada como na Equação 3.36:

$$LSC = 1 - \frac{1}{n^2} \sum_{i=1}^{n} |LS(x_i)|, \tag{3.36}$$

onde $|LS(x_i)|$ corresponde à cardinalidade do conjunto local do objeto x_i.

3.2.4 Medidas de rede

Nessa categoria de medidas, o conjunto de dados é representado como um grafo, preservando as distâncias ou similaridades entre os objetos originais. Nesse grafo, os vértices correspondem aos objetos, conectados por arestas ponderadas pela distância entre os objetos. As distâncias entre pares de pontos, medidas pela distância de Gower, são armazenadas em uma matriz (LORENA et al., 2019). O processo inclui a poda das arestas entre objetos de classes distintas. Para as medidas de rede, considera-se $G = (V, E)$ como o grafo obtido por esse processo, sendo que $|V| = n$ e $0 \leq |E| \leq \frac{n(n-1)}{2}$, o i-ésimo vértice do grafo é denotado como v_i, e a aresta entre dois vértices v_i e v_j é denotada como e_{ij}.

Densidade média da rede ($Density$)

Para o cálculo de $Density$ considera-se o número de arestas retidas no grafo construído a partir do conjunto de dados normalizado pelo número máximo de arestas entre n pares de dados, como definido pela Equação 3.37:

$$Density = 1 - \frac{2|E|}{n(n-1)}, \tag{3.37}$$

Valores baixos para essa medida indicam regiões densas de pontos conectados da mesma classe, correspondendo a uma baixa complexidade.

Coeficiente de agrupamento ($ClsCoef$)

A medida $ClsCoef$ é calculada como a taxa do número de arestas de os vizinhos de um vértice v_i e o máximo número de arestas que poderia existir entre eles, como definido pela equação 3.38:

$$ClsCoef = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{2|e_{jk} : vj, vk \in N_i|}{k_i(k_i - 1)}, \tag{3.38}$$

onde $N_i = \{vj : eij \in E\}$ corresponde aos nós diretamente conectados a v_i e k_i corresponde ao tamanho de N_i. Conjuntos de dados mais simples, com regiões mais densas de conexões entre objetos da mesma classe, terão valores mais baixos para $ClsCoef$.

Índice de pontos centrais ($Hubs$)

$Hubs$ calcula a influência dos nós no grafo atribuindo um índice a cada vértice, com base em suas conexões com outros vértices e com base no número de conexões de seus vizinhos. Na equação 3.39:

$$Hubs = 1 - \frac{1}{n}\sum_{i=1}^{n} hub(v_i), \tag{3.39}$$

$hub(v_i)$ é obtido como o autovetor principal de A^tA, onde A corresponde à matriz adjacente do grafo. Em conjuntos de dados com alta sobreposição de classes os vértices tenderão a ser menos conectados a vizinhos fortes, elevando o valor de $Hubs$.

3.2.5 Medidas de dimensionalidade

As medidas de dimensionalidade dão uma indicação da esparsidade dos dados com base na dimensionalidade do conjunto de dados.

Número médio de pontos por dimensão ($T2$)

Essa dimensão reflete a esparsidade do conjunto de dados calculando a razão entre a dimensionalidade e o conjunto de objetos do conjunto de dados, como na equação 3.40:

$$T2 = \frac{m}{n}. \tag{3.40}$$

onde m corresponde ao número de pontos e n corresponde ao número de dimensões. Conjuntos de dados menos esparsos indicam problemas mais simples.

Número médio de pontos por dimensões PCA ($T3$)

$T3$ estima a esparsidade do conjunto de dados calculando a quantidade média de pontos por componente PCA (PCA - Análise de Componentes Principais) necessário para representar 95% da variabilidade dos dados (m'), como representado na equação 3.41:

$$T3 = \frac{m'}{n}. \tag{3.41}$$

Taxa de dimensões PCA em relação às originais ($T4$)

$T4$ estima a proporção de dimensões relevantes para o dataset, e é calculada como na equação 3.42:

$$T4 = \frac{m'}{m}, \tag{3.42}$$

indicando que, quanto maior o valor de $T4$, mais variáveis são necessárias para descrever a variabilidade dos dados, e mais complexo o conjunto de dados.

3.2.6 Medidas de desbalanceamento de classes

Nessa categorias são apresentadas medidas que procuram capturar diferenças significativas no número de objetos das classes, que indicam problemas mais complexos.

Entropia das proporções de classe ($C1$)

Usada para estimar o desbalanceamento de classes, essa medida pode ser obtida pela estimativa da entropia normalizada da distribuição dos tamanhos das classes, e é calculada como na equação 3.43 (LORENA et al., 2012):

$$C1 = -\frac{1}{log(n_c)} \sum_{i=1}^{n_c} p_{c_i} log(p_{c_i}), \tag{3.43}$$

onde p_{c_i} corresponde à proporção de objetos da classe i e n_c corresponde ao número de classes. Quanto maior o valor de $C1$ mais balanceadas as classes (LORENA et al., 2019, p.16).

Taxa de desbalanceamento ($C2$)

$C2$ estima o balanceamento de classes pela equação 3.44:

$$C2 = 1 - \frac{1}{IR}, \tag{3.44}$$

onde IR é obtido pela equação 3.45:

$$IR = \frac{n_c - 1}{n_c} \sum_{i=1}^{n_c} \frac{n_{c_i}}{n - n_{c_i}}, \tag{3.45}$$

onde n_{c_i} corresponde ao número de objetos na classe i.

3.3 Dimensões de qualidade dos dados

Jayawardene, Sadiq e Indulska (2015) consolidam a definição e agrupam 127 dimensões da qualidade de dados, reunidas em produções dos contextos acadêmico e de negócios. Dependentes da definição adotada para dado, seja como representação de fatos, objetos ou pessoas (Liebenau, 1990), ou como matéria-prima para a informação (Wang, 1998), as dimensões são então categorizadas sob a perspectiva declarativa ou sob a perspectiva de uso. Na perspectiva declarativa são agrupadas dimensões da qualidade intrínseca que em si mesmas explicam os dados, tais como as impostas por metadados, por padrões de esquema ou por regras de negócio. Sob a perspectiva de uso estão as dimensões cuja avaliação é dependente do usuário, tais como aquelas relacionadas a eficiência e efetividade da criação do dado, e as relacionadas a usabilidade. As dimensões são ainda classificadas quanto à granularidade em que são aplicáveis: ao elemento de dado (atributo de uma entidade), ao registro de dado (coleção de atributos que formam uma entidade), ou ao objeto informacional (coleção de registros).

Os atributos das dimensões identificadas por Jayawardene, Sadiq e Indulska (2015) são, finalmente, utilizados para construir oito agrupamentos, posteriormente identificados por Completude, Disponibilidade e Acessibilidade, Atualidade, Acurácia, Validade, Usabilidade e Interpretabilidade, Confiabilidade e Credibilidade, e Consistência. Embora detalhadas, as categorias de dimensões de qualidade de dados precisam ter sua aplicabilidade a análises de Mineração de Dados verificada, e nessa direção a pesquisa de Berti-Equille (2007) contribui demonstrando o efeito de variações de atualidade, acurácia, incompletude e consistência sobre a mineração de regras de associação.

Já Hair et al. (2014) apontam discrepância (*outliers*) e incompletude (*missing values*) como problemas comuns na Análise Multivariada de Dados, descrevendo os procedimentos de identificação e tratamento desses problemas de qualidade em conjuntos de dados.

Para o interesse da presente pesquisa serão considerados apenas discrepância e incompletude, apontados por Hair et al. (2014) como problemas da qualidade de dados, representados por Jayawardene, Sadiq e Indulska (2015) pelos agrupamentos Completude e Validade. Juntas, essas duas dimensões fundamentais da qualidade de dados respondem por 30 das 127 dimensões tratadas por Jayawardene, Sadiq e Indulska (2015).

Os conceitos de discrepância e incompletude são melhor discutidos nas seções seguintes.

3.3.1 Discrepância

O conceito de discrepância (do Inglês *outlier*) pode ser associado tanto a um objeto com características diferentes da maioria dos objetos do conjunto de dados, como a um valor distintamente diferente da maioria das observações para o mesmo atributo, sem que esse objeto ou valor sejam considerados erros ou ruídos (TAN; STEINBACH; KUMAR, 2018).

A ocorrência de discrepâncias pode ser relacionada ao comportamento não usual dos processos geradores dos dados, refletindo características anormais dos sistemas e entidades que impactam esses processos. Por essa razão valores discrepantes também são chamados de anomalias, discordâncias, desvios ou anormalidades pelas literaturas de Mineração de Dados e de Estatística. A identificação, o tratamento ou mesmo a análise mais aprofundada de valores discrepantes ocorrem em passos diferentes do processo KDD: no passo de pré-processamento valores discrepantes podem ser detectados e, dependendo da análise em vista, serem tratados. Já no passo de mineração, valores discrepantes são o interesse da atividade de detecção de anomalias (AGGARWAL, 2015).

Embora haja consenso sobre a definição do que é um valor discrepante, a literatura parece não convergir quanto à sua classificação. Silva, Peres e Boscarioli (2016, p. 43) incluem valores discrepantes e os erros de medida sob o guarda-chuva de valores ruidosos, ao passo que Ferrari e Silva (2017, p. 27) classificam discrepância e violação de domínio como tipos de inconsistência. Já Tan, Steinbach e Kumar (2018, p. 41) consideram a discrepância e a inconsistência como diferentes tipos de problemas com os dados.

3.3.1.1 Métodos de detecção de valores discrepantes

A detecção de valores discrepantes se baseia na definição de um padrão do que são objetos ou pontos normais. Os valores discrepantes são, então, aqueles pontos ou objetos que não se encaixam no padrão de normalidade, em um grau maior ou menor, que é medido por uma pontuação de discrepância (*outlier score*). A pontuação de discrepância pode assumir uma escala real, indicando uma tendência maior ou menor de um ponto ser considerado discrepante, ou um valor binário, indicando se o ponto é ou não discrepante.

Aggarwal (2015) apresenta alguns modelos de análise de valores discrepantes:

- Análise de valores extremos - Sob a perspectiva da análise de valores extremos pontos discrepantes são pontos periféricos, correspondendo às caudas em uma distribuição probabilística;
- Modelos de agrupamento - Nessa perspectiva discrepantes são os pontos isolados dos agrupamentos de pontos;

- Modelos baseados em distância - Nesse modelo são considerados discrepantes os pontos com as maiores distâncias obtidas pelo algoritmo k-NN (k vizinhos mais próximos);
- Modelos baseados em densidade - Nesses modelos a pontuação de discrepância é definida tomando como base a densidade de pontos no entorno de um ponto;
- Modelos probabilísticos - Nessa categoria se calcula a probabilidade de um ponto ser gerado por um determinado modelo, cujos parâmetros são estimados a partir do conjunto de dados. Pontos discrepantes são aqueles com menor probabilidade de terem sido gerados pelo modelo;
- Modelos teórico-informáticos - Nesse modelo os pontos discrepantes são aqueles que adicionam complexidade ao código mínimo necessário para descrever a distribuição de um conjunto de dados.

A presente pesquisa adotou a definição de valores discrepantes como sendo os valores extremos em conjuntos de dados multivariados por ser de simples identificação, como mostra a próxima seção.

3.3.1.2 Detecção de discrepâncias pela análise de valores extremos

Pontos identificados como discrepantes pelo método de valores extremos correspondem àqueles pontos das caudas das distribuições probabilísticas uni e multidimensionais.

Em análises uni-variadas e para distribuições probabilísticas não simétricas a cauda superior é formada por todos os valores extremos superiores a um determinado limiar e a cauda inferior, pelos valores extremos inferiores. Considerando uma distribuição de densidade de pontos $f_X(x)$ as caudas serão definidas como as regiões extremas da distribuição para as quais $f_X(x) < \theta$, sendo θ um limiar de densidade de pontos definido pelo usuário. Numa distribuição bimodal, por exemplo, podem ocorrer regiões de pontos com distribuições inferiores a θ que podem ser considerados discrepantes, mas não são considerados valores extremos e, por isso, não são considerados na análise de valores extremos.

Embora em uma distribuição normal as áreas dentro das caudas superior e inferior representem a probabilidade acumulada dessas regiões extremas, o conceito de limiar de densidade (θ) é que define as caudas. Para uma distribuição normal adota-se como regra que pontos com valores absolutos de *z-score* maiores que 3 são considerados discrepantes, sendo a área interna das caudas inferior a 0,01%.

Em conjuntos de dados multivariados, os valores extremos serão aqueles com uma densidade de probabilidade inferior a um limiar. Assume-se como premissa que os pontos estão localizados em uma distribuição de probabilidades unimodal e que os pontos extremos serão os mais distantes do centro, em todas as direções. Pontos de interesse serão

aqueles cuja distância de Mahalanobis para a média do conjunto de dados for maior que um limiar (AGGARWAL, 2015) ou aqueles cujo valor de χ^2 seja maior que um limiar (FERRARI; SILVA, 2017, p.282).

A distância de Mahalanobis é uma medida de distância para conjuntos de dados multidimensionais em que cada objeto x_i é convertido em um escalar usando a seguinte métrica (Eq. 3.46):

$$r_i^2 = (x_i - \bar{x})^T \sum^{-1} (x_i - \bar{x}), \tag{3.46}$$

onde $\bar{x}$ corresponde ao vetor de valores médios dos atributos e $\sum$ corresponde à matriz de covariância do conjunto de dados.

Uma proposta para a identificação de valores extremos usando a distância de Mahalanobis é usar o valor de χ^2 com d graus de liberdade como limiar, a partir do qual valores extremos seriam identificados (AGGARWAL, 2015, p.243). No entanto, a distância de Mahalanobis é sensível ao efeito de mascaramento, que ocorre quando um agrupamento de valores discrepantes atrai a média aritmética do conjunto de dados e infla a matriz de covariância em sua direção (ROUSSEEUW; ZOMEREN, 1990).

Ye e Chen (2001) apresentam o método de detecção de anomalias baseada em valores extremos para bases de dados multivariadas, usando χ^2 como medida de distância. Nesse método, são considerados pontos extremos aqueles para os quais $\chi^2 > \bar{X}^2 + 3S_x^2$, sendo $\bar{X}^2$ a média dos atributos e S_x^2 o seu desvio padrão.

3.3.1.3 Contabilizando discrepâncias

A contabilização das ocorrências de valores discrepantes se dará por meio da dimensão Validade, seguindo a definição de Jayawardene, Sadiq e Indulska (2015), apresentada na Seção 3.3. A validade de um conjunto p pode ser definida como a taxa de dados válidos em relação ao total de dados do conjunto, podendo ser representada como (Equação 3.47):

$$V_p = 1 - \left(\frac{n}{N}\right), \tag{3.47}$$

sendo N o número de objetos do conjunto de dados p e n o número de objetos de dados extremos. Para a identificação dos objetos extremos utilizou-se a distância de Mahalanobis adotando-se χ^2 com d graus de liberdade como limiar, sendo d o número de atributos do conjunto de dados (AGGARWAL, 2015, p.243).

3.3.2 Incompletude

Incompletude pode ser compreendida como a medida que computa a taxa de valores ausentes ou nulos em um conjunto de dados. Um valor ausente é a falta de uma

medição em um conjunto de dados, cuja ausência prejudica a análise do fenômeno sendo estudado. A ausência de dados é comum não apenas nas Ciências Sociais mas em outras ciências e em situações cotidianas, sendo a inferência dos valores ausentes uma prática em que os critérios nem sempre são bem analisados e nem suas consequências, discutidas (MCKNIGHT et al., 2007; HAIR et al., 2014).

Um valor ausente pode ocorrer aleatoriamente no conjunto de dados em análise, estar relacionado ao comportamento dos participantes de um estudo, ao projeto do estudo ou à interação dos participantes com o projeto de estudo, ou ainda pode ocorrer devido a uma representação incorreta da entidade modelada, e também pode ocorrer porque o valor não é conhecido no momento da inserção do dado (MCKNIGHT et al., 2007; COUGO, 2013; HAIR et al., 2014).

Embora o interesse da presente pesquisa na ocorrência de valores ausentes esteja em contabilizá-los como uma dimensão da qualidade dos dados, os métodos de imputação ou extrapolação de valores ausentes passam a interessar na medida que interferem no cálculo de dimensões da complexidade dos dados e da qualidade das classificações.

3.3.2.1 Riscos dos valores ausentes

McKnight et al. (2007) defendem que os valores ausentes afetam cada uma das três etapas do método científico: a observação, a construção de inferências causais e a sua generalização.

Valores nulos podem afetar a etapa de observação de algum fenômeno de maneiras diferentes, dependendo, por exemplo, de quantos indicadores ou itens são usados para medir um determinado conceito ou construto. Em situações em que um construto é medido por apenas um indicador não há outros itens que possam contribuir para a substituição de valores ausentes, tornando-se a imputação um problema de inferência estatística. Problemas de ausência de valores, nesse caso, afetarão a confiabilidade da medida, o que pode vir a comprometer a validade do construto. Embora o efeito da ausência de valores tenda a ser menor em construtos medidos por vários itens, ainda assim há riscos de a validade do construto ser afetada, especialmente se os valores nulos ocorrerem em objetos significativos para uma amostra.

A validade interna de um modelo pode ser definida como o quanto uma relação entre dois construtos pode ser suportada pelas evidências. A ocorrência de valores ausentes afeta uma pesquisa em sua capacidade de ser reprodutível e em sua capacidade de generalização, duas facetas da validade interna. Um conjunto de dados com muitas ausências pode representar uma quantidade menor e enviesada de objetos, prejudicando a estimativa das relações modeladas, o que pode comprometer a validade interna do modelo.

Os efeitos dos valores ausentes em conjuntos de dados são agravados nos procedi-

mentos estatísticos de análise dos dados. A abordagem de eliminar objetos com valores nulos de um conjunto de dados reduz o tamanho da amostra e, como consequência, reduz também a capacidade de se detectar parâmetros significantes em correlações ou em análises comparativas, isto é, poder estatístico. Uma diminuição do poder estatístico implica aumento no risco do erro Tipo II (falha em rejeitar a hipótese nula quando essa é falsa). Outro efeito negativo dos valores ausentes em um conjunto de dados é notado na magnitude do efeito estatístico, no relacionamento entre dois construtos: valores ausentes podem diminuir a confiabilidade das medidas, diminuindo o poder estatístico. Ainda na Estatística, valores nulos podem afetar a distribuição dos dados e dos erros, com efeitos sobre testes paramétricos que pressuponham distribuições específicas.

Finalmente, a ausência de dados tem consequências para a generalização causal. Quando, por exemplo, se presume que uma amostra reduzida devido a valores ausentes ainda reflita a população de um estudo, as generalizações das descobertas refletirão os subgrupos presentes na amostra e não todos os subgrupos da população de interesse (MCKNIGHT et al., 2007).

A tabela 2 apresenta uma síntese das consequências dos dados ausentes nas três etapas do método científico.

Tabela 2 – Consequências dos dados ausentes. Setas para a direita significam como "leva a", setas para baixo significam "menor", setas para cima significam "maior". Adaptado de McKnight et al. (2007).

Tipo de consequência	Tipo de valor ausente	Aspecto do estudo afetado
Mensuração (confiabilidade, validade do construto)	Mensuração (itens, medidas únicas, múltiplas medidas do construto com dados ausentes)	↓ conjunto de itens → ↑ variância do erro → ↓ confiabilidade da medida ↓ informação → representação incompleta dos conceitos ↓ validade da medida
Confiabilidade e validade dos resultados do estudo (validade interna)	Seleção da amostra	Diferenças nas características dos grupos → viés de seleção → amostra pouco representativa → ↓ validade interna
	Análise dos dados (tamanho das amostras)	↓ poder estatístico e violação das suposições estatísticas → ↓ validade das conclusões estatísticas
Generalização dos resultados	Um ou mais dos tipos anteriores	Um ou mais dos problemas anteriores → dificuldade com a inferência estatística e com a interpretação dos resultados → base de conhecimento não acurada → recomendações mal informadas ou enganosas

3.3.2.2 Classificando valores ausentes

Categorizar os valores ausentes é uma estratégia adotada pela Estatística para facilitar a comunicação e contribuir para a medição de efeitos e soluções possíveis para cada classe de ausência. A categorização mais amplamente adotada nessa área de conhecimento é a que se baseia em Rubin (1976). Essa categorização é probabilística e considera as variáveis ou atributos que apresentam os valores ausentes, as variáveis ou atributos associados e o mecanismo hipotético subjacente aos dados ausentes, apresentando três classes de ausências, cada uma com um grau de impacto que os dados ausentes podem exercer sobre as análises estatísticas (MCKNIGHT et al., 2007; DAVEY et al., 2009; HAIR et al., 2014; FERRARI; SILVA, 2017).

O conjunto de dados ilustrado na Tabela 3, composto por três atributos, A, B e C, e por dez objetos, será utilizado para a explicação das classes de ausências propostas por Rubin (1976). O atributo A representa a variável independente, e os atributos B e C, as variáveis dependentes.

Tabela 3 – Conjunto de dados hipotético para apresentação das categorias de dados ausentes. Adaptado de McKnight et al. (2007).

Objeto	A	B	C
1	3	25	28
2	2	22	19
3	4	23	26
4	5	27	32
5	1	15	16
6	3	16	20
7	7	22	25
8	8	28	26
9	9	30	35
10	5	26	31

Supondo que valores ausentes ocorram apenas para a variável independente C, pode-se representar o conjunto de dados como na Tabela 4.

Na Tabela 4, as colunas D_A, D_B e D_C representam indicadores binários da ocorrência (1) ou não (0) de valores ausentes, formando a matriz R. Essa matriz contém apenas indicadores 0 quando não ocorrerem valores ausentes nos dados originais. As colunas MV_A, MV_B e MV_C compõem a matriz Y. A matriz Y pode ser subdividida na matriz de valores observados (não ausentes), Y_{obs}, e na matriz hipotética de valores ausentes, Y_{miss}.

Como pode ser visto na Figura 2, uma observação com valor ausente será classificada como MCAR (do inglês *Missing Completely At Random*) quando sua probabilidade de ocorrência não estiver relacionada a quaisquer valores observados ou não observados, ou $Pr(r|y_{obs}, y_{miss}) = Pr(r)$ (DAVEY et al., 2009). Na classe MCAR, qualquer obser-

Tabela 4 – Conjunto de dados hipotético. As células em côr preta indicam os valores ausentes. Adaptado de McKnight et al. (2007).

Dados originais				Dados com ausências (Y)			Valores *dummy* (R)		
Objeto	A	B	C	MV_A	MV_B	MV_C	D_A	D_B	D_C
1	3	25	28	3	25		0	0	1
2	2	22	19	2	22	19	0	0	0
3	4	23	26	4	23	26	0	0	0
4	5	27	32	5	27	32	0	0	0
5	1	15	16	1	15	16	0	0	0
6	3	16	20	3	16	20	0	0	0
7	7	22	25	7	22	25	0	0	0
8	8	28	26	8	28		0	0	1
9	9	30	35	9	30	35	0	0	0
10	5	26	31	5	26	31	0	0	0

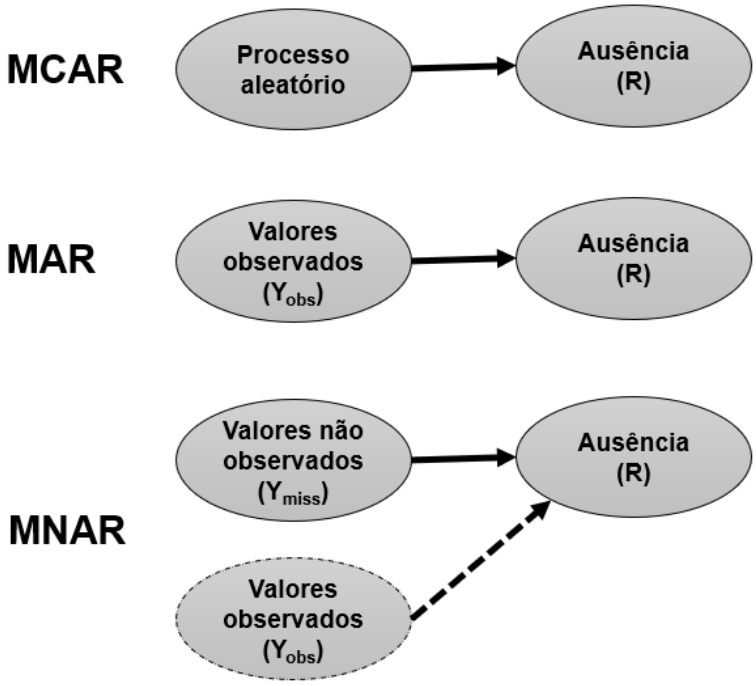

Figura 2 – Categorização de ausências proposta por Rubin (1976): MCAR (*Missing Completely At Random*), MAR (*Missing At Random*) e MNAR (*Missing Not At Random*). Fonte: McKnight et al. (2007).

vação terá a mesma probabilidade de conter valor ausente, uma vez que não existe um relacionamento sistemático entre R, Y_{obs} e Y_{miss}.

Por outro lado, uma ausência será classificada como MAR (do inglês *Missing At Random*) quando a probabilidade de sua ocorrência depender apenas das variáveis observadas, ou medidas, e nada das variáveis não medidas. Nesse caso, se existe um processo sistemático subjacente à ocorrência de valores ausentes, ele é governado por fatores para os quais não há dados (dados ausentes). Assim, MAR reflete um padrão de ausência relacionado a Y_{obs}, mas não a Y_{miss}, ou $Pr(r|y_{obs}, y_{miss}) = Pr(r|y_{obs})$ (MCKNIGHT et al., 2007; DAVEY et al., 2009).

Finalmente, em situações em que a probabilidade de uma observação ausente for atribuída unicamente a variáveis não observadas, a ausência é classificada como MNAR (do inglês *Missing Not At Random*). Em casos MNAR a ausência de valores não é igualmente provável entre os diferentes valores da variável, ou $Pr(r|y_{obs}, y_{miss}) = Pr(r|y_{miss})$, embora o processo subjacente à ocorrência dos valores ausentes não possa ser medido. É possível também que o relacionamento entre R e Y_{miss} envolva Y_{obs}.

A distinção entre MAR e MNAR requer suposições não testáveis a respeito da natureza das observações ausentes e sobre o processo subjacente, dificultando, portanto, uma diferenciação prática entre as duas classes de ausência (MCKNIGHT et al., 2007).

3.3.2.3 Um processo de tratamento de valores ausentes

O tratamento de valores ausentes é parte do passo de pré-processamento dos dados, dentro do processo de Descoberta de Conhecimento em Bases de Dados (KDD). Dentro desse passo, na etapa de limpeza dos dados ocorrem as atividades de imputação de valores ausentes, remoção de ruídos e correção de inconsistências, podendo ocorrer a eliminação de uma instância com ocorrências de valores ausentes (HAIR et al., 2014; SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017).

Hair et al. (2014) delineia um processo de quatro passos para identificar e corrigir valores ausentes em conjuntos de dados, considerando o tipo, a extensão e a aleatoriedade das ocorrências de valores ausentes para, então, apresentar o método de imputação mais apropriado (Figura 3). Imputação é o nome que se dá ao processo que atribui um valor a uma ocorrência de ausência de dado, permitindo que análises estatísticas sejam feitas no conjunto de dados. Embora a imputação produza um conjunto de dados com menos valores ausentes, esse conjunto de dados não deve ser considerado um conjunto de dados real (PROJECT, 2009).

Os tipos de valores ausentes diferenciam as ausências sob controle do pesquisador daquelas cujos impactos e causas são desconhecidos. Quando os parâmetros que governam o processo de ausência de dados não estão relacionados aos parâmetros a serem estimados na pesquisa, os valores ausentes podem ser ser considerados ignoráveis e não necessitam de ações corretivas específicas.

Em casos de ausências ignoráveis, ou o processo que produz os valores ausentes está operando aleatoriamente ou está considerado na técnica de análise de dados utilizada. Em outras palavras, são considerados ignoráveis dados ausentes classificados como MCAR, que não têm efeito sistemático na estimativa dos parâmetros, e aqueles ausentes classificados como MAR em que o processo de geração de ausências é conhecido e os valores ausentes podem ser repostos com base em dados observados do conjunto (RUBIN, 1976; MCKNIGHT et al., 2007; HAIR et al., 2014). Por outro lado, processos de geração de ausências não são ignoráveis quando não podem ser modelados com os dados dispo-

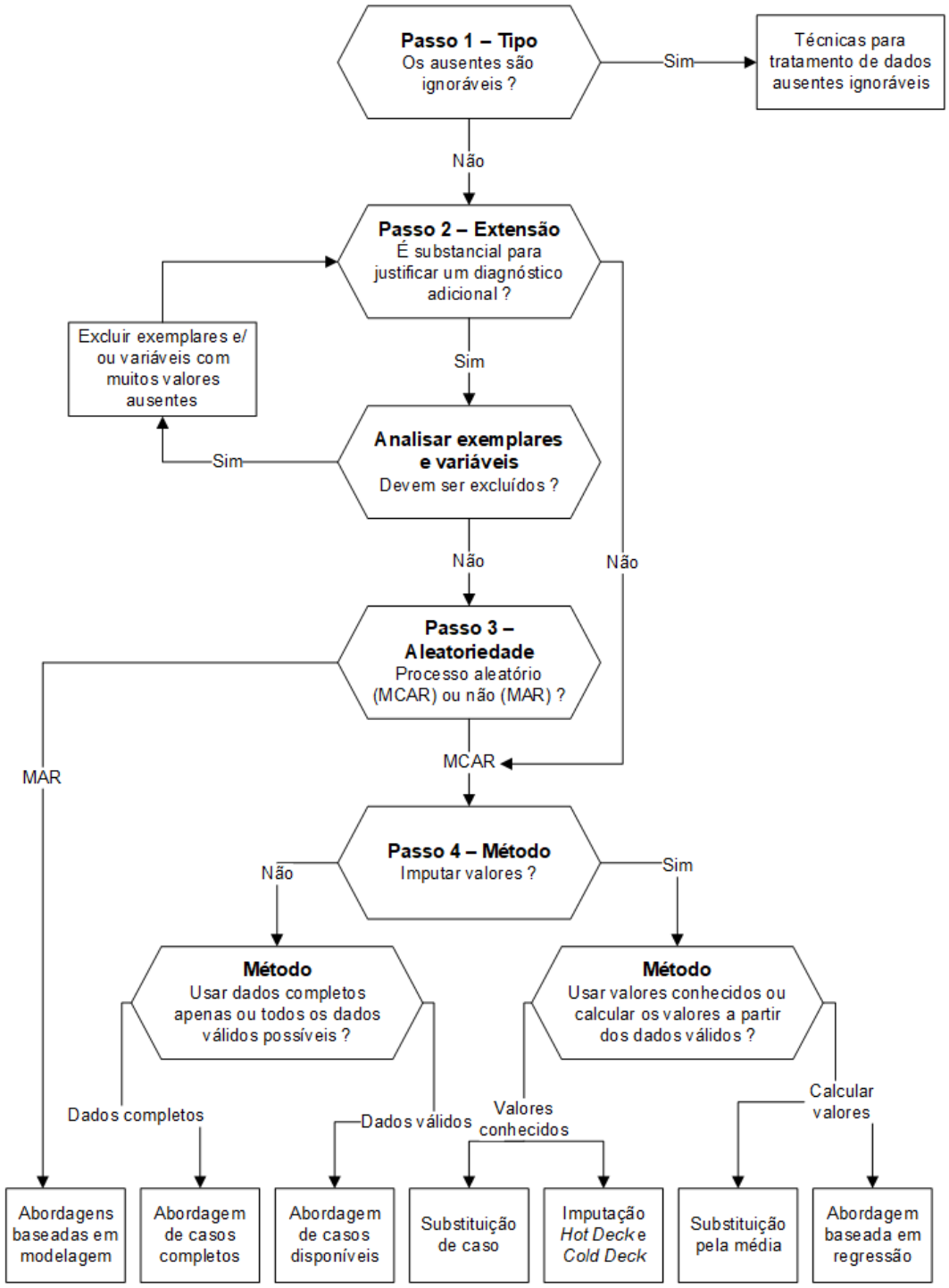

Figura 3 – Processo de identificação e correção de valores ausentes. Fonte: Hair et al. (2014).

níveis, gerando um impacto desconhecido na estimativa dos parâmetros e nas conclusões estatísticas (MCKNIGHT et al., 2007).

Exemplos de processos de ausências de dados ignoráveis incluem: a) dados de uma população não incluídos em um processo de amostragem aleatória, b) dados não coletados em um instrumento de coleta por não se aplicarem a parte dos respondentes, e c) dados ausentes devido a uma restrição conhecida na amostra analisada. Por outro lado, processos

de dados não ignoráveis podem incluir processos conhecidos para o pesquisador, como erros nos procedimentos de coleta dos dados ou dados não preenchidos devido à morte de um participante de uma pesquisa em andamento, e processos desconhecidos, como indisposição do respondente em preencher questões controversas ou sensíveis (HAIR et al., 2014).

O passo inicial do processo de identificação e correção de valores ausentes (Passo 1, na Figura 3) consiste na identificação do tipo do processo que gerou o dado ausente em análise naquele instante, podendo ausências ignoráveis e não ignoráveis ocorrer simultaneamente no mesmo conjunto de dados. Se a ausência for identificada como não ignorável, o próximo passo (Passo 2, na Figura 3) é analisar se a extensão e o impacto das ausências nas variáveis, nos casos e no conjunto de dados como um todo, são baixos o suficiente para não afetar os resultados das análises.

Hair et al. (2014) apresentam a discussão sobre o que pode ser considerado como uma extensão baixa o suficiente para não afetar os resultados de uma análise, propondo um método e duas regras. O método proposto para analisar a extensão da ocorrência de ausências consiste na tabulação de a) o percentual de variáveis com dados ausentes em cada caso ou ocorrência, b) o número de casos com valores ausentes em cada variável, e c) o número de casos sem a ocorrência de dados ausentes para qualquer variável. Como exemplo de aplicação, a tabulação entre os dados de "a" e "b" permitiria ao analista perceber se há concentração de valores ausentes em variáveis específicas, enquanto o dado obtido em "c" indicaria o tamanho da amostra caso nenhuma ação corretiva seja tomada.

A pergunta central do Passo 2 procura encontrar um desequilíbrio entre o viés que pode ocorrer caso não se tome qualquer ação corretiva para as ocorrências de valores ausentes e o viés que pode ocorrer na análise caso valores ausentes sejam substituídos. Se a extensão dos valores ausentes for considerada aceitável (poucas ocorrências produzidas por um processo aleatório), podem ser adotadas técnicas de imputação sem o risco de viés nas análises (Passo 4, na Figura 3). Ao contrário, se a extensão da ocorrência de valores ausentes não for considerada baixa, a aleatoriedade do processo subjacente à geração de valores ausentes deve ser analisada (Passo 3, na Figura 3). Como parâmetros para avaliar a extensão da ocorrência de valores nulos, Hair et al. (2014) propoem:

- um percentual de dez porcento de ocorrências de valores nulos em um objeto do conjunto de dados é considerado baixo, exceto se a causa subjacente não for aleatória, e
- caso não se opte pela imputação, o número de objetos sem a ocorrência de valores ausentes deve ser suficiente para a análise desejada.

A exclusão de objetos ou de variáveis com valores ausentes é uma opção para

o pesquisador, desde que se considere que uma amostra reduzida pode comprometer a confiabilidade da análise. Novamente, Hair et al. (2014) apresentam parâmetros para avaliar se a exclusão de objetos ou variáveis deve ocorrer:

- variáveis ou casos com cinquenta por cento ou mais valores ausentes devem ser excluídos, mas sob risco de inviabilizar a análise;
- variáveis com quinze por cento de valores ausentes são candidatas à exclusão, mas até mesmo ausências de vinte ou trinta por cento podem ser corrigidas;
- a exclusão de variáveis ou objetos se justifica se produzir um conjunto de dados mais apto às análises;
- ocorrências de valores ausentes em variáveis dependentes devem ser excluídas para evitar viés no relacionamento com as variáveis independentes;
- na exclusão de uma variável, outra variável que represente a excluída, preferencialmente altamente correlacionada, deve ser mantida, e
- as análises devem ser feitas com e sem os casos ausentes tratados para que se identifique o efeito do tratamento.

No Passo 3 do processo apresentado na Figura 3 se avalia a presença ou não de aleatoriedade no processo subjacente à ocorrência de valores ausentes. Em casos de conjuntos de dados pequenos, testes visuais podem ser feitos para esse fim, ao passo que em conjuntos de dados com muitas variáveis e objetos, testes estatísticos podem indicar se há um padrão não aleatório por detrás das ocorrências de valores ausentes. Testes estatísticos podem analisar e comparar se o padrão de aleatoriedade se assemelha ao padrão esperado de um processo aleatório de geração de valores ausentes, podendo classificá-lo como MCAR ou MAR. Em qualquer um dos casos já opções para imputação de valores.

No Passo 4 do processo (Figura 3) se escolhe o método de imputação de dados. Métodos para ausências MAR tratam os valores ausentes como parte da análise, mas os métodos de imputação para ausências do tipo MCAR substituem os valores ausentes tentando identificar relacionamentos nos dados observados que possam contribuir na imputação dos valores ausentes. (HAIR et al., 2014). A Tabela 5 menciona alguns dos procedimentos de tratamento de valores ausentes, tanto para as do tipo MAR quanto MCAR:

3.3.2.4 Contabilizando valores ausentes

A ocorrência completamente aleatória de valores ausentes (MCAR) é mais desejável se comparada à ocorrência da mesma quantidade de valores ausentes concentrada em

Tabela 5 – Alguns procedimentos de tratamento de ausências. Adaptado de Hair et al. (2014).

Procedimento	Descrição	Indicação
	Imputação usando apenas dados válidos	
Dados completos	Considera na análise apenas os objetos com dados completos	Amostras grandes, fortes relacionamentos entre as variáveis, poucas ausências
Todos os dados disponíveis	Considera apenas os dados presentes como representativos da amostra completa em cálculos de medidas de distribuição e de relação	Poucas ausências, relacionamentos moderados entre as variáveis
	Imputação com um valor conhecido	
Substituição do objeto	Substitui o objeto por um outro de dentro ou de fora da amostra, caso exista, selecionado pela similaridade com o objeto substituído	Quando existem objetos adicionais à amostra e que possuam similaridade com os objetos da amostra
Imputação *hot deck* e *cold deck*	Na imputação *hot deck* o atributo com valor ausente é substituído pelo valor do mesmo atributo de um outro objeto da amostra que seja similar ao objeto que contenha o valor ausente. Na imputação *cold deck*, no entanto, o objeto similar provém de uma fonte externa, fora da amostra.	Requer outras medidas sobre as quais a similaridade seja calculada
	Imputação por um valor calculado	
Imputação pela média	O valor ausente é calculado pela média dos valores do mesmo atributo dos demais objetos válidos. Pode ser ajustado para usar a média por subgrupos dentro da amostra.	Poucos valores ausentes e relacionamento forte entre as variáveis
Imputação por modelos preditivos	Utiliza modelos preditivos baseados em regressão ou classificação para estimar valores para as ausências.	Graus moderados ou altos de valores ausentes, relacionamentos suficientes entre as variáveis para não impactar a generalização
	Métodos para ausências do tipo MAR	
Baseados em modelos	Aplicáveis a ausências do tipo MAR, incorporam os valores ausentes às análises, seja estimando os valores ausentes por um processo explicitamente projetado, ou considerando os valores ausentes uma parte do processo de análise multivariada	Muitos dados ausentes e necessidade de uso de um método que introduza menos viés e que garanta generalização

algumas variáveis ou em variáveis mais significativas para a análise (HAIR et al., 2014, p.40). Assim, medir a completude de um conjunto de dados usando uma medida simples para computação de valores ausentes pode não ser tão adequada.

No entanto, a opção do pesquisador é utilizar uma medida de completude que supõe a aleatoriedade do processo de geração das ocorrências de valores ausentes (MCAR) dos conjuntos de dados analisados (Subseção 4.2.3). A completude de um conjunto p pode ser definida como a taxa de dados existentes em relação ao que deveria existir, independente da causa, podendo ser representada como (Equação 3.48)(BLAKE; MANGIAMELI, 2011):

$$C_p = 1 - \left(\frac{n}{N(1+A)}\right), \tag{3.48}$$

sendo N o número de objetos do conjunto de dados p, A o número de atributos e n o número de ocorrências de valores ausentes.

3.3.3 Sensibilidades dos algoritmos de classificação da pesquisa a valores ausentes e discrepantes

Como evidenciam o passo de pré-processamento do processo KDD e todas as preocupações destacadas nos itens anteriores, a ausência de valores e a ocorrência de valores discrepantes tendem a reduzir a qualidade das análises de um conjunto de dados. No entanto, resta saber a forma como esses problemas afetam os algoritmos de classificação de dados adotados nas presente pesquisa como indicadores da qualidade da classificação.

Dentre os algoritmos de classificação estão aqueles chamados de *instance-based learning algorithms*. Esses algoritmos utilizam-se de objetos similares para efetuar a classificação. O algoritmo k-NN (*k-Nearest Neighbors*) classifica um objeto considerando a classe da maioria dos seus k vizinhos mais próximos, com base em uma medida de distância ou similaridade entre os atributos dos objetos comparados (SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017). Uma vez que o cálculo de similaridade ou dissimilaridade baseia-se em operações de comparação ou subtração, valores ausentes não permitem essas operações. Quando a quantidade de vizinhos com os quais o objeto é comparado é muito pequena, a ocorrência de valores discrepantes nos atributos dos vizinhos pode afetar o resultado da classificação (TAN; STEINBACH; KUMAR, 2018).

Algoritmos de árvore de decisão se baseiam em uma estrutura de árvore, construída com base nos dados de treinamento, para classificar novos objetos. Nessa árvore os nós internos correspondem a testes de atributos, os ramos correspondem a resultados dos testes e as folhas representam as classes. A escolha mais apropriada para um nó é determinada por um algoritmo de indução (SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017). C4.5 é um algoritmo de indução de árvores de decisão que utiliza a informação como critério para decisão do atributo de quebra da árvore. Na indução da árvore de decisão a ausência de valores é considerada na construção dos nós, seja tratando os valores nulos como um possível ramo, seja distribuindo as ocorrências com nulos entre

os ramos respeitando a distribuição dos dados por meio de pesos. No momento da classificação uma ocorrência com um dado ausente faz com que cada um dos ramos do nó correspondente ao atributo seja testado (QUINLAN, 2014). Por outro lado, valores discrepantes podem conduzir o processo de indução a um sobreajuste da árvore aos dados, requerendo algoritmos que tratem o sobreajuste (TAN; STEINBACH; KUMAR, 2018).

Outro algoritmo baseado em árvore é o CART (*Classification And Regression Trees*), que utiliza o índice Gini para decisão do atributo de quebra da árvore. O algoritmo CART ignora as ocorrências com valores ausentes na medição da qualidade de uma quebra, e usa quebras substitutas (*surrogate splits*) para determinar como lidar com valores ausentes na etapa de teste da classificação (FEELDERS, 1999). Da mesma forma como no algoritmo C4.5, árvores induzidas pelo algoritmo CART requerem tratamento para evitar o sobreajuste. Quanto ao efeito de valores discrepantes sobre o algoritmo CART, a literatura defende que o algoritmo CART não é significantemente afetado por valores discrepantes em variáveis independentes, mas é afetado localmente por valores discrepantes presentes nas variáveis dependentes (HäRDLE; SIMAR, 2015, p.552) (NISBET; ELDER; MINER, 2009, p.161).

Random Forests é uma classe de métodos de classificação baseados em árvore que aplica o método de *bootstrapping* ao conjunto de dados para diminuir a variância da predição. As predições de múltiplas árvores são combinadas para apresentar o modelo de classificação. *Random Forests* são resistentes a ruídos e discrepâncias de variáveis independentes, pelo fato de aplicarem a normalização pelo método de encaixotamento às variáveis (AGGARWAL, 2015, p.381). Na definição original de *Random Forests* proposta por Breiman e Cutler (2004), valores ausentes são tratados por uma de duas formas, ambas por imputação. Implementações de *Random Forests* usando árvores de decisão C4.5 ao invés de CART, adotam a abordagem do algoritmo C4.5 para lidar com valores ausentes.

Support Vector Machines (SVM) são algoritmos de classificação baseados em teorias de aprendizado estatístico, desenvolvidos inicialmente para conjuntos de dados numéricos e binários (duas classes), embora adaptações no conjunto de dados e nos algoritmos permitam o uso de dados categóricos e problemas multi-classes (TAN; STEINBACH; KUMAR, 2018, p.276). Algoritmos SVM buscam construir uma função de um hiperplano que maximize as margens que separam duas classes linearmente separáveis ou que separam a maioria dos pontos de duas classes não linearmente separáveis. Para a construção desse hiperplano se utiliza um subconjunto dos dados de treinamento, cujos pontos são chamados de vetores de suporte (*support vectors*). Uma vez que os algoritmos SVM focam extensivamente na região de separação entre duas classes, essa categoria de algoritmos têm um bom comportamento de generalização diante de conjuntos de dados com valores discrepantes (AGGARWAL, 2015, p.321). SVM não são tolerantes a valores ausentes, requerendo tratamento prévio dos dados (HäRDLE; SIMAR, 2015).

Redes Neurais Artificiais é uma classe de algoritmos de classificação que simulam o sistema nervoso humano para computar. O aprendizado de uma rede neural ocorre pelo ajuste dos pesos dos nós (neurônios) de entrada que compõem a função computacional da rede neural, tomando como base os dados de treinamento, que funcionam como estímulos externos. O arranjo entre os neurônios de uma rede neural é chamado de arquitetura, sendo Perceptron de camada única a arquitetura mais básica. Na rede Perceptron há duas camadas de nós: os nós de entrada, cuja quantidade corresponde à dimensionalidade do conjunto de dados, e um nó de saída. Cada nó de entrada recebe um valor de entrada numérico e o transmite para o nó de saída, que executa uma função matemática quando recebe uma entrada. A rede neural Perceptron de uma camada executa classificações binárias. Os pesos dos nós de entrada são combinados para aprender uma função linear, chamada função de ativação, que gera uma saída entre -1,+1.

As redes neurais podem implementar modelos mais complexos pela adição de mais camadas, mais neurônios, e funções de ativação arbitrárias, requerendo mais épocas de treinamento para a aprendizagem dos pesos dos diferente nós. Embora as redes neurais possam aproximar qualquer função de classificação, a complexidade da configuração e a dificuldade de definir adequadamente a quantidade de épocas de treinamento da rede podem levar a um sobreajuste da rede aos dados de treinamento. A rede neural pode algumas vezes ser sensível a dados discrepantes nos dados de treinamento (AGGARWAL, 2015, p.330). Objetos com valores ausentes nos dados de treinamento devem ser retirados ou ter seus valores imputados (TAN; STEINBACH; KUMAR, 2018, p.255).

Tabela 6 – Comportamento dos algoritmos de classificação utilizados na pesquisa diante de valores ausentes e discrepantes.

Algoritmo	Tipo	Valores ausentes	Valores discrepantes
C4.5	Baseado em árvores	Lida com valores ausentes	Lida com valores discrepantes, requerendo algoritmos para evitar o sobreajuste
Classification And Regression Trees (CART)	Baseado em árvores	Lida com valores ausentes	Não significantemente afetado por valores discrepantes em variáveis independentes, mas afetado localmente por valores discrepantes nas variáveis dependentes
Random Forests (RF)	Baseado em árvores	Lida com valores ausentes	Lida com valores discrepantes

A Tabela 6 sumariza o comportamento dos algoritmos de classificação utilizados na presente pesquisa diante de valores ausentes e discrepantes. A pesquisa optou inicialmente

por não utilizar algoritmos que requeressem tratamento prévio de valores ausentes, por imputação de dados ou por deleção de objetos, e de valores discrepantes.

4 Procedimentos Metodológicos

A proposta de trabalho da presente pesquisa foi propor um modelo de caminho (modelo estrutural e modelo de mensuração) que ilustre de maneira simples a hipótese de relacionamento entre os conceitos de qualidade de dados, qualidade da classificação e complexidade dos dados. Definidos o modelo estrutural e o modelo de mensuração, os tamanhos de amostra mínimo e recomendado para o cálculo do modelo estrutural foram calculados e os dados, obtidos. Uma vez descritos, os dados foram submetidos ao cálculo do modelo estrutural e às validações dos resultados.

Por se tratar de um método de análise multivariada exploratória, o modelo estrutural inicialmente proposto poderia ser modificado para um modelo mais complexo visando a uma melhor adequação do modelo aos fatos observados (HAIR et al., 2016, pp.4,229).

Nas seções seguintes são apresentados os detalhes metodológicos dos experimentos e os resultados obtidos.

4.1 Proposição dos modelos estrutural e de mensuração

A literatura oferece subsídios para a proposição de um modelo estrutural em que a qualidade da classificação de conjuntos de dados é afetada pela complexidade e pela qualidade dos dados analisados. Como qualidade e complexidade de dados e qualidade da classificação são conceitos cuja observação não é direta, passam a ser chamados de construtos ou variáveis latentes, sendo medidos indiretamente por meio de indicadores que manifestam-se como dimensões de qualidade e de complexidade (HAIR et al., 2016). O relacionamento entre os construtos fundamenta-se em fatos tais como o efeito que a presença de valores discrepantes exerce sobre dimensões de complexidade dependentes da variância, como por exemplo, medidas de sobreposição, podendo afetar o resultado de classificadores como Árvores de Decisão e *Support Vector Machines*, como discutido na subseção 3.3.3.

As múltiplas perspectivas que a literatura oferece ao tentar definir Qualidade de Dados e a ausência de uma definição formal para Complexidade de Dados conduziram a tarefa de identificação dos construtos mais para a caracterização do que para a nomeação. Assim, optou-se por uma adaptação da definição da complexidade de Kolgomorov (1965) para Complexidade dos Dados, como apresentada em Boschetti (2008), e por seguir a ideia de Jayawardene, Sadiq e Indulska (2015) ao definir Qualidade de Dados. A presente pesquisa propôs definir os construtos da seguinte maneira:

- **Complexidade dos Dados** - Esforço necessário para descrever um conjunto de dados. Quanto maior o esforço, mais complexos são os dados;
- **Qualidade dos Dados** - Fidelidade com que os dados representam pessoas, objetos, eventos ou conceitos. Quanto maior a qualidade, maior a proximidade entre a representação e o objeto ou fato representado;
- **Qualidade da Classificação** - Eficácia com que objetos de teste são classificados corretamente. Quanto maior a qualidade, maior a proximidade dos objetos de teste de suas verdadeiras classes.

4.1.1 Modelo estrutural

A relação entre os construtos pode ser representada como na Figura 4. Nota-se que o construto Qualidade dos Dados exerce efeito direto sobre Qualidade da Classificação e também sobre Complexidade dos Dados, sugerindo um efeito mediador da Complexidade dos Dados sobre a Qualidade da Classificação.

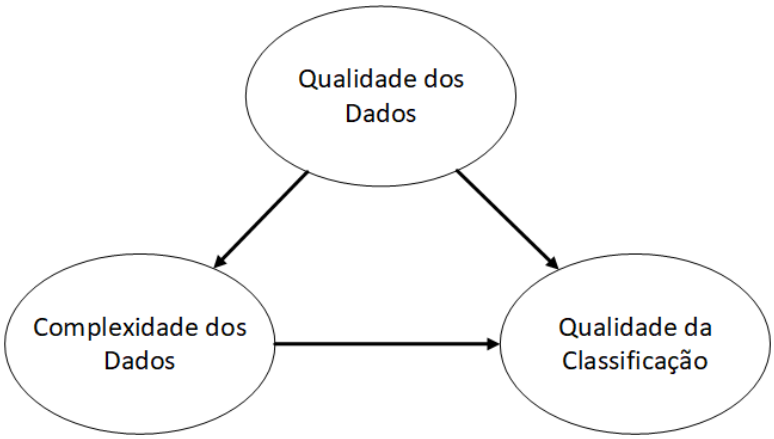

Figura 4 – Modelo de relação dos construtos.

4.1.2 Modelo de mensuração

O modelo de mensuração dos construtos, apresentado na Figura 5, fundamentou-se teoricamente pela literatura apresentada nas Subseções 3.2 e 3.3. Compuseram os indicadores dos construtos as dimensões de qualidade apresentadas nas Seções 3.3.1.3 e 3.3.2.4 (Indicadores **a** e **b** na Figura 5), as dimensões de complexidade de dados apresentadas na Tabela 1 (Indicadores **p**, **q** e **r** na Figura 5), bem como os resultados dos classificadores listados na Tabela 6 (Indicadores **x**, **y** e **z** na Figura 5).

Cabe nesse ponto ponderar que as dimensões da Qualidade dos Dados adotadas capturaram expressões diferentes de problemas causadores de falta de qualidade em conjuntos de dados. Da discussão apresentada na Subseção 3.3 é possível depreender que

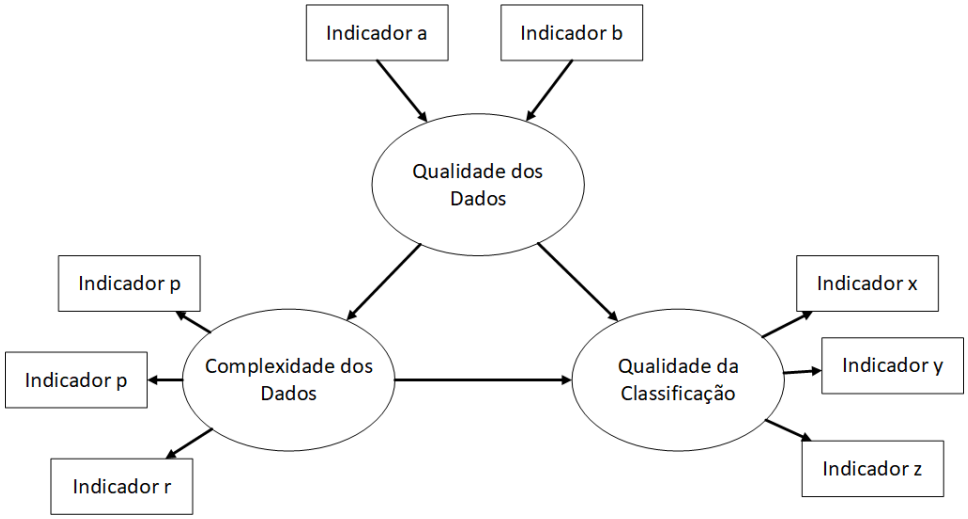

Figura 5 – Modelo de mensuração dos construtos, apresentando indicadores formativos e reflexivos.

validade e completude cobrem o domínio de conteúdo da qualidade dos dados de interesse para o presente estudo, a saber, aspectos da qualidade que afetam a classificação de dados.

Quanto à Complexidade dos Dados, a definição do papel reflexivo ou formativo das dimensões adotadas na presente pesquisa baseou-se nos esforços de Lorena et al. (2019) em apresentar diferentes medidas que afetam a classificação de dados, não se descartando a possibilidade de haver sobreposição de indicadores, isto é, indicadores medindo os mesmos fenômenos. Assim, os indicadores que representaram a Complexidade dos Dados exerceram um caráter reflexivo e entre eles esperou-se encontrar alguma correlação. Finalmente, os indicadores adotados para medir o construto Qualidade da Classificação expressaram-se como o resultado de diferentes algoritmos para a mesma tarefa (Tabela 6). Como são causados pelo mesmo construto, esperou-se desses indicadores que apresentassem alta correlação entre si (HAIR et al., 2016). Na Figura 5 é possível identificar a relação de reflexão ou formação dos indicadores com os construtos.

Uma vez que a relação entre os construtos Qualidade da Classificação, Complexidade dos Dados e Qualidade dos Dados estava suficientemente clara e fundamentada, passou-se à coleta dos dados.

4.2 Conjunto de dados experimental

A ocorrência de valores ausentes e discrepantes em um conjunto de dados é uma possibilidade prevista no processo de Descoberta de Conhecimento em Bases de Dados (KDD). O passo de pré-processamento desse processo inclui atividades de tratamento de valores ausentes, tais como a imputação ou a eliminação de objetos, e de tratamento de valores discrepantes, tais como o encaixotamento (*binning*), o agrupamento e a aproximação (SILVA; PERES; BOSCARIOLI, 2016; FERRARI; SILVA, 2017).

Uma hipótese da presente pesquisa é que a qualidade dos dados afeta sua complexidade e, por consequência, a qualidade das classificações. Foi de interesse da pesquisa conhecer o efeito de dados contendo valores discrepantes e ausentes não submetidos ao pré-processamento sobre os resultados de algoritmos de classificação. Assim, os experimentos executados na presente pesquisa previram a constituição de um conjunto de dados que contivesse resultados de algoritmos de classificação para conjuntos de dados contendo valores discrepantes e valores ausentes não tratados.

A constituição e a quantidade de objetos do conjunto de dados utilizado nos experimentos são discutidos nas subseções seguintes.

4.2.1 Quantidade de objetos do conjunto de dados experimental

O tamanho do conjunto de dados experimental tem relação direta com poder estatístico, isto é, com a probabilidade de se rejeitar uma hipótese nula quando ela é falsa. Ao se adotar um conjunto de dados com valores ausentes, diminui-se o poder estatístico e corre-se o risco de comprometer uma distribuição de dados que se assume ter com os dados completos (FACELI et al., 2000; MCKNIGHT et al., 2007).

Para o cálculo do tamanho do conjunto de dados amostral adequado à Modelagem de Equações Estruturais, assumiram-se os seguintes valores apresentados na tabela 7.

Tabela 7 – Critérios para cálculo do tamanho mínimo do conjunto de dados amostral. Fonte: Hair et al. (2016, pp.20-22).

Critério	Valor
Tamanho mínimo do efeito que se deseja detectar como significante	0,3 (médio)
Poder estatístico desejado	80%
Nível de significância	5%

O número de variáveis observadas foi calculado da seguinte maneira:

- **22** dimensões de complexidade aplicáveis à tarefa de classificação, Tabela 1
- **2** dimensões de qualidade de dados, Subseções 3.3.1.3 e 3.3.2.4
- **3** algoritmos de classificação, Tabela 6

Dado o número inicial de **27** variáveis observadas e de **3** variáveis latentes do modelo estrutural (Figura 4), o tamanho amostral mínimo necessário para detectar-se o efeito, calculado com base na tabela apresentada por Hair et al. (2016, p.21), foi de cerca de **59** objetos, e na calculadora de Soper (2017) foi de **67** objetos. A pesquisa adotou a quantidade de **67** objetos como tamanho amostral mínimo.

4.2.2 Ferramental para a formação do conjunto de dados experimental

Para a coleta dos dados experimentais foram utilizadas as seguintes ferramentas:

- Suite R, versão 4.0.3 (R Core Team, 2018)
- Pacote Rcommander, versão 2.7-1 (FOX, 2005)
- Pacote ECoL, versão 0.4.0 (GARCIA et al., 2020)
- Pacote StatMeasures, versão 1.0 (JAIN, 2015)
- Pacote OutlierDetection, versão 0.1.1 (TIWARI; KASHIKAR, 2019)

4.2.3 Constituição do conjunto de dados experimental

A abordagem adotada para a obtenção do conjunto de dados utilizado nos experimentos foi a de construir um espaço de medidas de complexidade e qualidade de dados para problemas de classificação no qual os atributos fossem medidas de complexidade de dados, medidas de qualidade de dados e medidas de desempenho dos classificadores.

Em outras palavras, o conjunto de dados experimental constituiu-se de metadados coletados de bases de dados reais, em que cada linha contém metadados de uma base de dados real e cada coluna representa um metadado da base de dados. A Figura 6 exemplifica o conjunto de dados experimental.

x_{ij}

	ID	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	$i_{...}$	i_n
		Data Quality dimensions				Data Complexity dimensions				Classification Quality dimensions			
	datasetName	**DQ1**	**DQ2**	**...**	**DQn**	**DC1**	**DC2**	**...**	**DCn**	**CQ1**	**CQ2**	**...**	**CQn**
1	DSName1												
2	DSName2												
3	DSName3												
...	....												
n	DSNameN												

$\vec{x}_i$

Figura 6 – Diagrama representativo do conjunto de dados experimental.

4.2.3.1 Atributos

Por atributos do conjunto de dados foram selecionadas: **a)** as medidas das dimensões de complexidade aplicáveis à tarefa de classificação listadas por Lorena et al. (2019) e apresentadas na Tabela 1, **b)** as medidas das dimensões validade e completude calculadas como descrito as Subseções 3.3.1.3 e 3.3.2.4, e **c)** os valores de AUC (*Area Under the ROC Curve*) que representassem o desempenho dos algoritmos de classificação da Tabela 6. As

colunas com os índices i_1 a i_n da Figura 6 representam os atributos do conjunto de dados experimental.

O indicador AUC representa o grau de separabilidade, isto é, quanto um algoritmo é capaz de distinguir entre as classes, e sua medida varia entre [0,1]. Para o cálculo de AUC utilizou-se o pacote ROCR (SING et al., 2005), aplicado a cada um dos algoritmos de classificação da Tabela 6, utilizando-se o método de validação cruzada em k-pastas, para k=10. Para o cálculo das medidas das dimensões de complexidade utilizou-se o pacote ECol, desenvolvido na linguagem R por Garcia et al. (2020).

Para a identificação dos pontos extremos utilizou-se a função de Mahalanobis, implementada no pacote *stats* da linguagem R, adotando-se p=0,95 e χ^2 com d graus de liberdade como limiar, sendo d o número de atributos do conjunto de dados. Definiu-se a tolerância de 1e−20 como limite máximo para que a função *solve* assuma valores muito pequenos como zero.

4.2.3.2 Objetos

Os objetos para o conjunto de dados experimental foram obtidos do repositório OpenML (CASALICCHIO et al., 2017), tendo sido pesquisados pelos seguintes critérios:

- número de objetos **entre 100 e 3.000**
- número de atributos **entre 2 e 20**
- número de classes **igual a 2**
- número de valores ausentes **igual ou maior que 0**

A razão de terem sido adotados conjuntos de dados binários é porque a maioria das medidas de complexidade é definida para problemas de classificação binária apenas, embora problemas multi-classes possam ser decompostos em binários pela estratégia OVO (*One-vs-One*) (LORENA et al., 2019). A quantidade de atributos foi definida levando-se em consideração que a dimensionalidade do conjunto de dados afeta sua complexidade (HO; BASU, 2000; HO; BASU, 2002; LORENA; CARVALHO, 2004; SÁNCHEZ; MOLLINEDA; SOTOCA, 2007; GARCIA; CARVALHO; LORENA, 2015; ZUBEK; PLEWCZYNSKI, 2016; BARELLA et al., 2018). A quantidade de valores ausentes definida no critério de seleção visou à obtenção de conjuntos de dados mistos, isto é, com e sem valores ausentes. O intervalo para o número de instâncias foi definido arbitrariamente.

Aplicados os critérios, foram obtidos **178** conjuntos de dados do repositório OpenML, já descontados os conjuntos de dados eliminados devido a problemas com seus dados. Des-

ses 178 objetos coletaram-se os metadados para a construção do espaço de medidas de complexidade e qualidade de dados para problemas de classificação.

4.3 Pré-processamento

As atividades de pre-processamento recomendadas pelo processo de KDD, a saber, limpeza, redução, transformação e discretização, incluindo o tratamento de valores ausentes e de discrepâncias, foram suprimidas nas bases de dados analisadas, com o objetivo específico de manter presentes os problemas de complexidade e de qualidade desses conjuntos.

No entanto, alguns indicadores possuem uma relação inversa com seus construtos, isto é, medem inversamente seus efeitos, conforme mostrado na Tabela 1. Exemplo dessa relação é o indicador F1v, taxa máxima discriminante vetor-direcional de Fisher, em que valores maiores indicam uma menor complexidade do conjunto de dados (LORENA et al., 2019, p.3). O pacote ECoL já trata esses indicadores apresentando os resultados em uma relação positiva.

5 Resultados e Discussões

5.1 Análise descritiva

As subseções seguintes apresentam a análise descritiva do conjunto de dados experimental. Para os resumos estatísticos utilizou-se o pacote *Rcommander* da linguagem R (FOX, 2005).

5.1.1 Estrutura do conjunto de dados

O conjunto de dados experimental é composto por metadados para a construção do espaço de medidas de complexidade e qualidade de dados para problemas de classificação, além de alguns atributos descritivos dos conjuntos de dados utilizados para a coleta de metadados. Os atributos do conjunto de dados experimental são (Tabela 8):

Tabela 8 – Atributos do conjunto de dados experimental, calculados conforme detalhado nas Subseções 3.2, 3.3.2.4, 3.3.1.3 e 3.3.3. No atributo datasetName, o domínio é qualquer nome.

Atributo	Descrição	Tipo	Domínio
datasetName	Nome do conjunto de dados	Categórico	*
MDAttributes	Quantidade de atributos do conjunto de dados	Discreto	[2,20]
MDElements	Quantidade de objetos no conjunto de dados	Discreto	[100, 3000]
DQMissingValues	Quantidade de valores ausentes no conjunto de dados	Discreto	≥ 0
DQOutliers	Quantidade de valores discrepantes no conjunto de dados	Discreto	≥ 0
DQCompleteness	Indicador de completude do conjunto de dados	Contínuo	[0,1]
DQValidity	Indicador de validade do conjunto de dados	Contínuo	[0,1]
B1	Entropia das proporções de classe	Contínuo	[0,1]
B2	Taxa de desbalanceamento	Contínuo	[0,1]
D1	Número médio de pontos por dimensão	Contínuo	≥ 0
D2	Número médio de pontos por dimensões PCA	Contínuo	≥ 0
D3	Taxa de dimensões PCA em relação às originais	Contínuo	[0,1]
F1	Taxa máxima discriminante de Fisher	Contínuo	[0,1]
F1v	Taxa máxima discriminante vetor-direcional de Fisher	Contínuo	[0,1]
F2	Volume da região de sobreposição	Contínuo	[0,1]
F3	Eficiência máxima individual do atributo	Contínuo	[0,1]
F4	Eficiência coletiva do atributo	Contínuo	[0,1]
L1	Soma do erro de distância por programação linear	Contínuo	[0,1]
L2	Taxa de erro de classificador linear	Contínuo	[0,1]
L3	Não-linearidade de um classificador linear	Contínuo	[0,1]
N1	Fração de pontos na fronteira da classe	Contínuo	[0,1]
N2	Taxa média de distância NN extra/intra classe	Contínuo	[0,1]
N3	Taxa de erro *leave-one-out* do classificador 1NN	Contínuo	[0,1]
N4	Não-linearidade do classificador 1-NN	Contínuo	[0,1]
N5	Fração de hiperesferas cobrindo os dados	Contínuo	[0,1]
N6	Cardinalidade média do conjunto local	Contínuo	[0,1]
G1	Densidade média da rede	Contínuo	[0,1]
G2	Coeficiente de agrupamento	Contínuo	[0,1]
G3	Indice de pontos centrais	Contínuo	[0,1]
C4.5	Indicador AUC do algoritmo de classificação C4.5	Contínuo	[0,1]
RF	Indicador AUC do algoritmo de classificação *Random Forests*	Contínuo	[0,1]
CART	Indicador AUC do algoritmo de classificação *Classification And Regression Trees*	Contínuo	[0,1]

A análise descritiva do conjunto de dados é detalhada nas próximas subseções, dividida por conjuntos de atributos.

5.1.2 Atributos descritivos

Para os metadados indicativos da quantidade de atributos (*MDAttributes*), de objetos (*MDElements*) e Dimensionalidade (*MDAttributes* * *MDElements*) algumas medidas de tendência central, de dispersão e de posição relativa dos dados são:

MDAttributes	MDElements	Dimensionalidade
Min. : 2.000	Min. : 100.0	Min. : 240
1st Qu.: 5.000	1st Qu.: 179.0	1st Qu.: 1122
Median : 7.000	Median : 329.5	Median : 2496
Mean : 7.876	Mean : 450.7	Mean : 3559
3rd Qu.:10.000	3rd Qu.: 555.5	3rd Qu.: 4940
Max. :19.000	Max. :2201.0	Max. :21440

Quanto à quantidade de elementos das bases de dados analisadas, nota-se que 75% das bases têm até 556 objetos e 10 atributos, e que a dimensionalidade média é de cerca de 2.500 valores (um valor ocorre em cada intersecção de uma linha com uma coluna).

5.1.3 Atributos de qualidade de dados

Quanto aos metadados de valores ausentes e de valores discrepantes, as seguintes medidas estatísticas foram obtidas:

DQMissingValues	DQOutliers	DQCompleteness	DQValidity
Min. : 0.00	Min. : 0.000	Min. :0.5488	Min. :0.9000
1st Qu.: 0.00	1st Qu.: 2.000	1st Qu.:1.0000	1st Qu.:0.9772
Median : 0.00	Median : 5.000	Median :1.0000	Median :0.9881
Mean : 35.48	Mean : 5.107	Mean :0.9925	Mean :0.9809
3rd Qu.: 0.00	3rd Qu.: 8.750	3rd Qu.:1.0000	3rd Qu.:0.9971
Max. :1368.00	Max. :18.000	Max. :1.0000	Max. :1.0000

	mean	**sd**	n
DQMissingValues	35.4831461	169.88175431	178
DQCompleteness	0.9924822	0.03984842	178
DQOutliers	5.1067416	4.20181015	178
DQValidity	0.9808865	0.02289886	178

A distribuição de frequências dos metadados indicadores da ocorrência de valores ausentes e discrepantes é apresentada na Figura 7.

Ao se compararem a quantidade de valores ausentes pela dimensão do conjunto de dados (atributos vezes objetos), considerando apenas os conjuntos de dados contendo ocorrências de valores ausentes, é possível perceber uma pequena correlação positiva entre a quantidade de valores ausentes e a dimensão do conjunto de dados. Nota-se também que 90% das ocorrências de valores ausentes encontram-se em conjuntos de dados com até 10.000 valores.

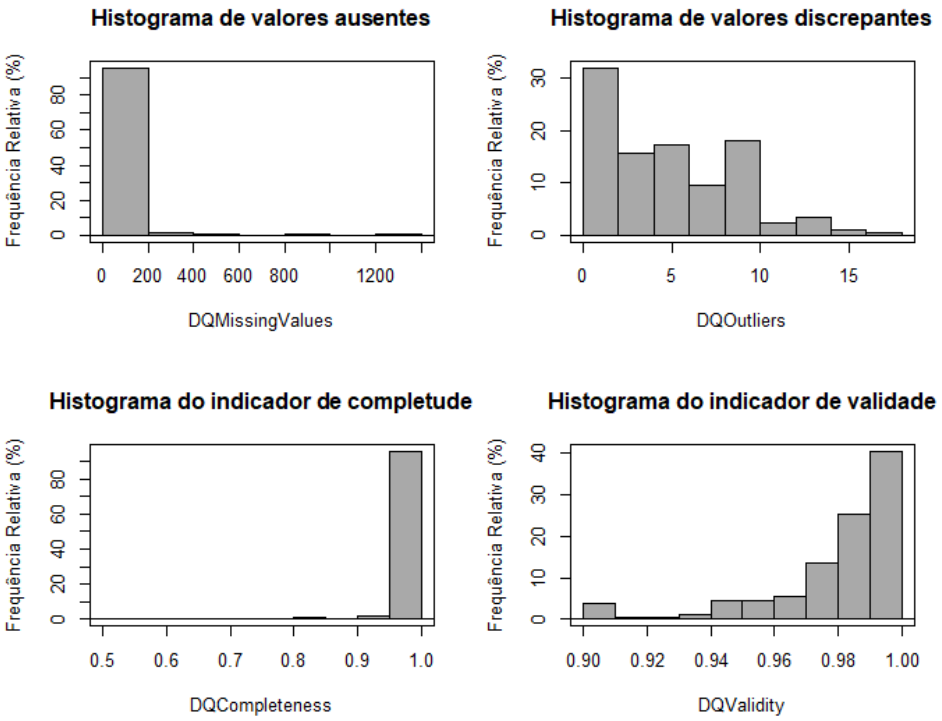

Figura 7 – Histograma dos metadados de valores ausentes e discrepantes.

Quanto às discrepâncias, nota-se uma maior correlação positiva entre a quantidade de valores discrepantes e a dimensionalidade do conjunto de dados analisado, 90% ocorrendo em conjuntos de dados com até cerca de 8.000 valores.

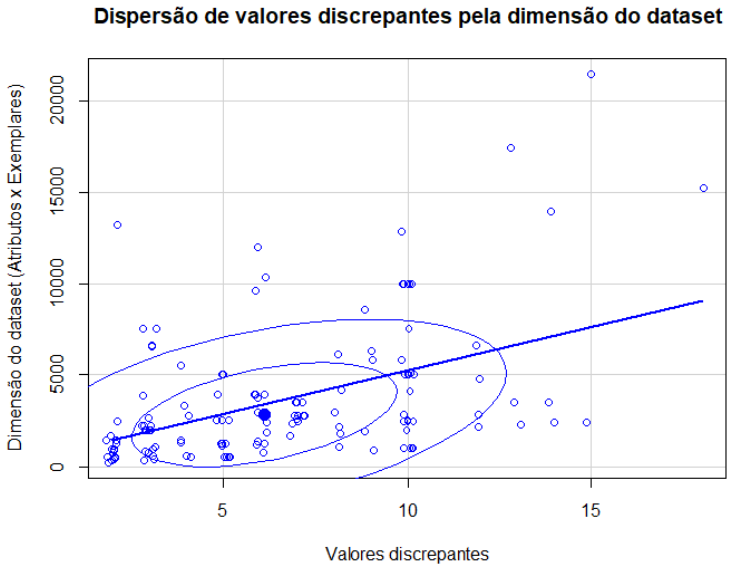

Figura 8 – Gráfico de dispersão entre a quantidade de valores discrepantes (>0) e a dimensão do conjunto dados, incluindo a linha de regressão e elipses indicando a concentração de 50% e 90% dos dados.

Para os indicadores de completude (DQCompleteness) e validade (DQValidity) os

valores de correlação de Pearson são:

	DQCompleteness	DQValidity
DQCompleteness	1.0000000	−0.0394095
DQValidity	−0.0394095	1.0000000

A baixa correlação entre os indicadores de completude e validade pode ser considerada aceitável, se for considerado que esses indicadores medem diferentes aspectos da qualidade dos dados.

Finalmente, a dispersão entre conjuntos de dados com ocorrências de valores ausentes e conjuntos de dados com ocorrências de valores discrepantes (Figura 9) revela que a maioria das ocorrências de valores discrepantes encontra-se em conjuntos de dados sem valores ausentes.

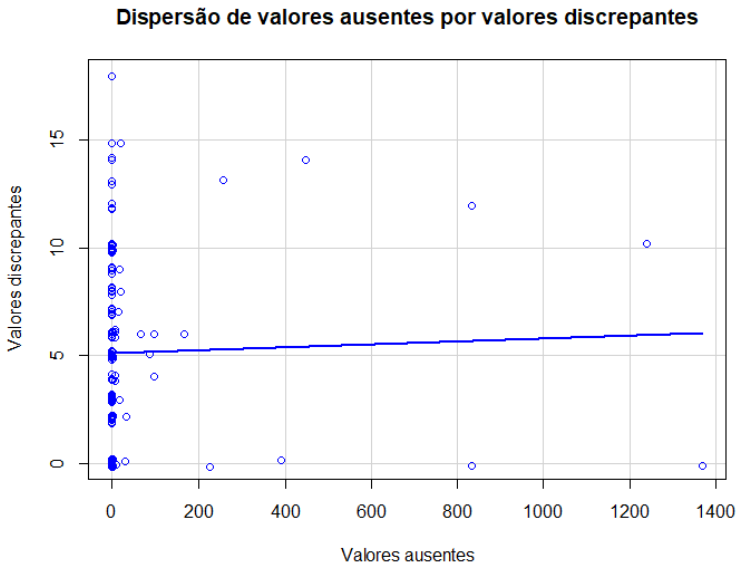

Figura 9 – Gráfico de dispersão entre a quantidade de valores ausentes e a quantidade de valores discrepantes.

5.1.4 Atributos de complexidade de dados

Para os indicadores de complexidade de dados, os valores da correlação de Pearson são:

	B1	B2	D1	D2	D3	F1	F1v	F2
B1	1.00000	0.97521	0.0379	−0.02650	−0.31391	0.14708	−0.2021	−0.3044
B2	0.97521	1.00000	0.0630	0.00508	−0.32053	0.14491	−0.2011	−0.3236
D1	0.03793	0.06297	1.0000	0.86243	−0.12947	0.00250	−0.2287	−0.2122
D2	−0.02650	0.00508	0.8624	1.00000	0.17644	−0.00831	−0.2044	−0.0823
D3	−0.31391	−0.32053	−0.1295	0.17644	1.00000	0.00665	0.1452	0.5913
F1	0.14708	0.14491	0.0025	−0.00831	0.00665	1.00000	0.5183	0.2159
F1v	−0.20215	−0.20105	−0.2287	−0.20443	0.14522	0.51833	1.0000	0.3202
F2	−0.30443	−0.32361	−0.2122	−0.08225	0.59128	0.21595	0.3202	1.0000
F3	−0.19565	−0.18027	−0.0187	0.05519	0.30361	0.61458	0.4674	0.4854
F4	−0.29122	−0.26861	−0.2878	−0.17030	0.34195	0.49052	0.5391	0.6004

G1	−0.40931	−0.41444	0.0476	0.11380	0.30543	0.55201	0.5497	0.4962
G2	−0.00284	0.01968	0.3011	0.39041	0.35611	0.32006	0.1195	0.2748
G3	−0.56671	−0.57299	−0.0758	−0.04342	0.07642	−0.29357	0.0958	0.0477
L1	−0.32671	−0.32446	−0.2067	−0.16032	0.20229	0.52070	0.8228	0.4249
L2	−0.38759	−0.40473	−0.1976	−0.15397	0.24933	0.52713	0.8838	0.4166
L3	−0.33603	−0.33702	−0.1151	−0.09864	0.23672	0.57157	0.8384	0.4174
N1	−0.35157	−0.33278	0.1086	0.16407	0.22690	0.52406	0.6597	0.3038
N2	−0.29529	−0.26052	0.0992	0.13124	0.12901	0.51693	0.4578	0.1187
N3	−0.30429	−0.29039	0.0717	0.05693	0.08742	0.50879	0.6748	0.2001
N4	−0.18696	−0.14792	0.0423	−0.02930	−0.01548	0.51362	0.5928	0.3130
N5	−0.35846	−0.34025	0.0954	0.16704	0.34875	0.52643	0.4968	0.3436
N6	−0.34227	−0.29429	−0.1362	−0.10476	0.21626	0.68354	0.4605	0.3615

	F3	F4	G1	G2	G3	L1	L2	L3	N1
B1	−0.1957	−0.2912	−0.4093	−0.00284	−0.5667	−0.327	−0.388	−0.3360	−0.352
B2	−0.1803	−0.2686	−0.4144	0.01968	−0.5730	−0.324	−0.405	−0.3370	−0.333
D1	−0.0187	−0.2878	0.0476	0.30110	−0.0758	−0.207	−0.198	−0.1151	0.109
D2	0.0552	−0.1703	0.1138	0.39041	−0.0434	−0.160	−0.154	−0.0986	0.164
D3	0.3036	0.3419	0.3054	0.35611	0.0764	0.202	0.249	0.2367	0.227
F1	0.6146	0.4905	0.5520	0.32006	−0.2936	0.521	0.527	0.5716	0.524
F1v	0.4674	0.5391	0.5497	0.11951	0.0958	0.823	0.884	0.8384	0.660
F2	0.4854	0.6004	0.4962	0.27480	0.0477	0.425	0.417	0.4174	0.304
F3	1.0000	0.8177	0.6446	0.38219	−0.0959	0.645	0.611	0.6284	0.609
F4	0.8177	1.0000	0.6064	0.31953	−0.0433	0.676	0.650	0.6391	0.570
G1	0.6446	0.6064	1.0000	0.36700	0.1684	0.690	0.724	0.7214	0.747
G2	0.3822	0.3195	0.3670	1.00000	−0.3992	0.144	0.174	0.2286	0.394
G3	−0.0959	−0.0433	0.1684	−0.39917	1.0000	0.189	0.185	0.0803	0.136
L1	0.6453	0.6756	0.6904	0.14401	0.1890	1.000	0.905	0.8190	0.718
L2	0.6111	0.6502	0.7237	0.17397	0.1852	0.905	1.000	0.9143	0.772
L3	0.6284	0.6391	0.7214	0.22863	0.0803	0.819	0.914	1.0000	0.733
N1	0.6093	0.5702	0.7465	0.39365	0.1355	0.718	0.772	0.7329	1.000
N2	0.4777	0.4283	0.4418	0.24338	−0.0231	0.506	0.538	0.5082	0.661
N3	0.5540	0.5179	0.6454	0.24986	0.1331	0.692	0.747	0.7387	0.889
N4	0.5472	0.5835	0.6639	0.19192	0.0337	0.663	0.651	0.7217	0.676
N5	0.5552	0.4984	0.6784	0.48523	0.0211	0.559	0.613	0.6182	0.770
N6	0.7390	0.6915	0.6594	0.27707	−0.0262	0.594	0.576	0.5752	0.601

	N2	N3	N4	N5	N6
B1	−0.2953	−0.3043	−0.1870	−0.3585	−0.3423
B2	−0.2605	−0.2904	−0.1479	−0.3402	−0.2943
D1	0.0992	0.0717	0.0423	0.0954	−0.1362
D2	0.1312	0.0569	−0.0293	0.1670	−0.1048
D3	0.1290	0.0874	−0.0155	0.3487	0.2163
F1	0.5169	0.5088	0.5136	0.5264	0.6835
F1v	0.4578	0.6748	0.5928	0.4968	0.4605
F2	0.1187	0.2001	0.3130	0.3436	0.3615
F3	0.4777	0.5540	0.5472	0.5552	0.7390
F4	0.4283	0.5179	0.5835	0.4984	0.6915
G1	0.4418	0.6454	0.6639	0.6784	0.6594
G2	0.2434	0.2499	0.1919	0.4852	0.2771
G3	−0.0231	0.1331	0.0337	0.0211	−0.0262
L1	0.5064	0.6917	0.6631	0.5595	0.5943
L2	0.5383	0.7469	0.6507	0.6129	0.5755
L3	0.5082	0.7387	0.7217	0.6182	0.5752
N1	0.6613	0.8887	0.6764	0.7700	0.6009
N2	1.0000	0.7738	0.3258	0.7849	0.5970
N3	0.7738	1.0000	0.6426	0.7856	0.5668
N4	0.3258	0.6426	1.0000	0.4296	0.5402
N5	0.7849	0.7856	0.4296	1.0000	0.6232
N6	0.5970	0.5668	0.5402	0.6232	1.0000

O valor do coeficiente alfa de Cronbach, que mede a confiabilidade interna de um construto reflexivo, bem como o ganho de confiabilidade em caso de exclusão de algum dos indicadores, são apresentados a seguir:

```
Alpha reliability =  0.8449
Standardized alpha =  0.8889

Reliability deleting each item in turn:
     Alpha Std.Alpha r(item, total)
B1   0.862     0.905         −0.3081
B2   0.872     0.905         −0.3244
D1   0.853     0.897         −0.0815
D2   0.847     0.895          0.0696
D3   0.848     0.890          0.3225
F1   0.836     0.879          0.6482
F1v  0.826     0.878          0.6552
F2   0.839     0.886          0.4870
F3   0.820     0.876          0.7786
F4   0.821     0.878          0.7322
G1   0.843     0.875          0.7830
G2   0.839     0.886          0.4452
G3   0.852     0.899         −0.0927
L1   0.836     0.876          0.7807
L2   0.829     0.875          0.7906
L3   0.830     0.875          0.7990
N1   0.822     0.874          0.7983
N2   0.834     0.880          0.5961
N3   0.830     0.875          0.7433
N4   0.834     0.879          0.6453
N5   0.822     0.876          0.7290
N6   0.836     0.878          0.6993
```

Um valor de alfa de Cronbach de 0,889 é considerado um valor bom de confiabilidade para o modelo de mensuração do construto de complexidade de dados. Embora esse resultado de confiabilidade possa ser melhorado pela exclusão de indicadores, optou-se pela manutenção de todos os indicadores por representarem diferentes medidas de complexidade, pertencentes a seis tipos de medidas de complexidade.

Para os indicadores de complexidade de dados, algumas medidas de tendência central, de dispersão e de posição relativa dos dados são apresentadas a seguir, agrupadas pela categoria das medidas.

Medidas de desbalanceamento de classes

Embora o indicador B1, entropia das proporções de classe, meça o desbalanceamento das classes, e tenha o seu maior valor para classes balanceadas (LORENA et al., 2019, p.16), a implementação desse indicador pelo pacote ECoL já calcula o valor complementar para manter uma relação direta entre o valor do indicador e a complexidade do dado (GARCIA et al., 2020). Por sua vez, o indicador B2, taxa de desbalanceamento, já mantém uma relação direta com o construto Complexidade do Dado, não necessitando ser ajustado.

A distribuição de frequências dos indicadores B1 e B2 pode ser vista na Figura 10.

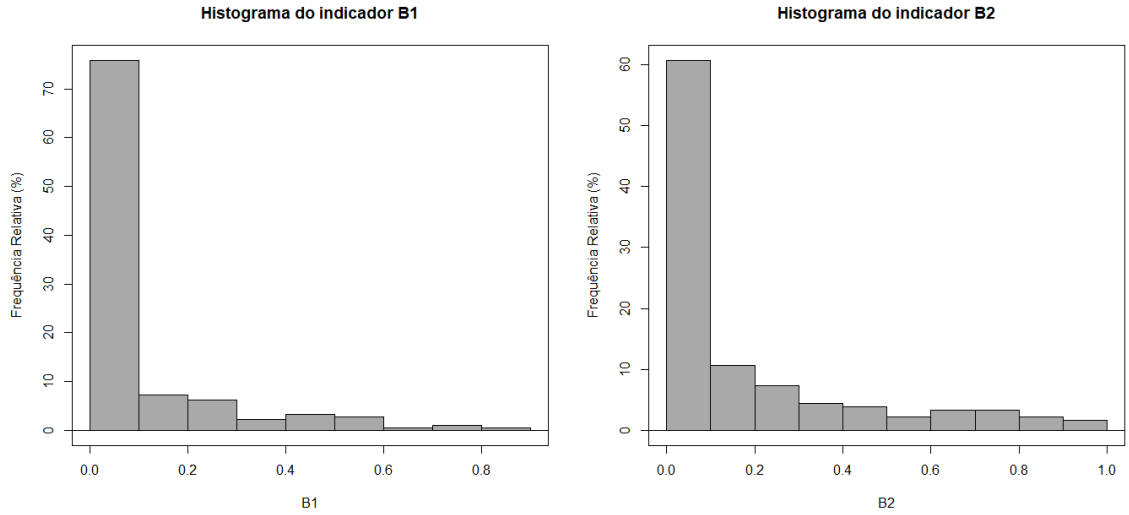

Figura 10 – Histograma dos indicadores B1 e B2.

Os valores de algumas medidas estatísticas desses indicadores para o conjunto de dados experimental são:

```
B1                  B2
Min.    :0.000000   Min.    :0.000000
1st Qu.:0.003137    1st Qu.:0.008654
Median :0.016743    Median :0.045195
Mean    :0.098462   Mean    :0.176373
3rd Qu.:0.098343    3rd Qu.:0.235088
Max.    :0.840650   Max.    :0.952408
```

Os valores do indicador B1, entropia das proporções de classe, interpretados considerando a inversão implementada no pacote ECoL, indicam que 75% dos conjuntos de dados analisados possuem um baixo desbalanceamento de classes, e os 25% restantes possuem uma taxa de desbalanceamento entre 0,098 e 0,84, numa escala de 0 a 1. Essa informação parece ser confirmada pelas medidas de posição relativa do indicador B2, taxa de desbalanceamento.

A alta correlação positiva entre os indicadores B1 e B2 pode ser observada abaixo:

```
    B1          B2
B1 1.0000000 0.9752104
B2 0.9752104 1.0000000
```

Medidas de dimensionalidade

As medidas de dimensionalidade indicam a esparsidade dos dados tomando como base os atributos dos conjuntos de dados. Nessas medidas, quanto maior o valor, mais atributos são necessárias para representar a variabilidade dos dados, e mais complexo é conjunto de dados.

O indicador D1, número médio de pontos por dimensão, reflete o esparsamento dos dados. Valores maiores para D1 indicam um problema menos complexo (LORENA et al., 2019, p.15). O pacote ECoL já retorna o valor numa relação positiva. O indicador D2, número médio de pontos por dimensões PCA, reflete o esparsamento dos dados entre atributos identificados pelo processo de Análise de Componentes Principais (PCA) como sendo responsáveis por 95% da variabilidade dos dados. Da mesma forma como no indicador D1, valores maiores para o indicador D2 representam problemas menos complexos, sendo ajustados pelo pacote ECoL. Finalmente, o indicador D3, taxa de dimensões PCA em relação às originais, mede a proporção de dimensões relevantes no conjunto de dados, guardando uma relação direta entre seu valor e a complexidade do conjunto de dados.

A distribuição de frequências dos indicadores D1, D2 e D3 pode ser vista na Figura 11.

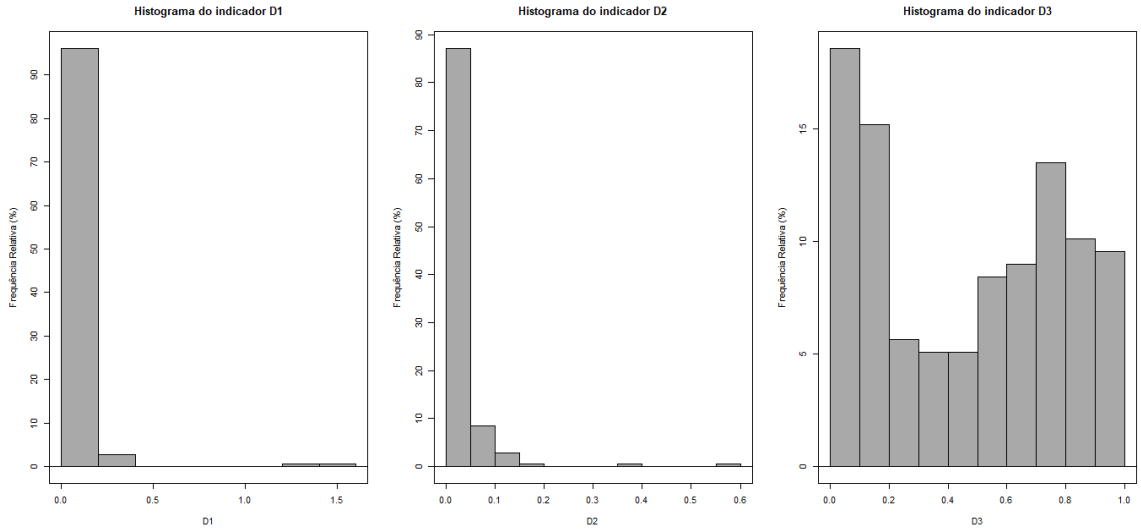

Figura 11 – Histograma dos indicadores D1, D2 e D3.

Algumas medidas estatísticas dos indicadores D1, D2 e D3 são apresentadas abaixo:

```
D1                  D2                  D3
Min.   :0.001363    Min.   :0.000500    Min.   :0.01471
1st Qu.:0.017500    1st Qu.:0.004532    1st Qu.:0.14286
Median :0.037344    Median :0.010000    Median :0.55051
Mean   :0.068094    Mean   :0.026849    Mean   :0.48783
3rd Qu.:0.072786    3rd Qu.:0.026466    3rd Qu.:0.80000
Max.   :1.474748    Max.   :0.583333    Max.   :1.00000
```

Os valores da dimensão D3 indicam que em 75% dos conjuntos de dados analisados a variabilidade dos dados pode ser explicada com até 55% dos atributos desses conjuntos.

A correlação entre os indicadores D1, D2 e D3 é apresentada abaixo:

```
      D1          D2          D3
D1  1.0000000  0.8624251 −0.1294680
D2  0.8624251  1.0000000  0.1764375
D3 −0.1294680  0.1764375  1.0000000
```

A baixa correlação negativa do indicador D3 com o indicador D1 não pôde ser explicada pela definição dessas medidas.

Medidas baseadas em atributos

As medidas baseadas em atributos procuram avaliar quão informativos os atributos do conjunto de dados são para separar as classes. Quanto menos informativos os atributos forem, mais complexo o problema.

A distribuição de frequências dos indicadores F1, F1v, F2, F3 e F4 pode ser vista na Figura 12.

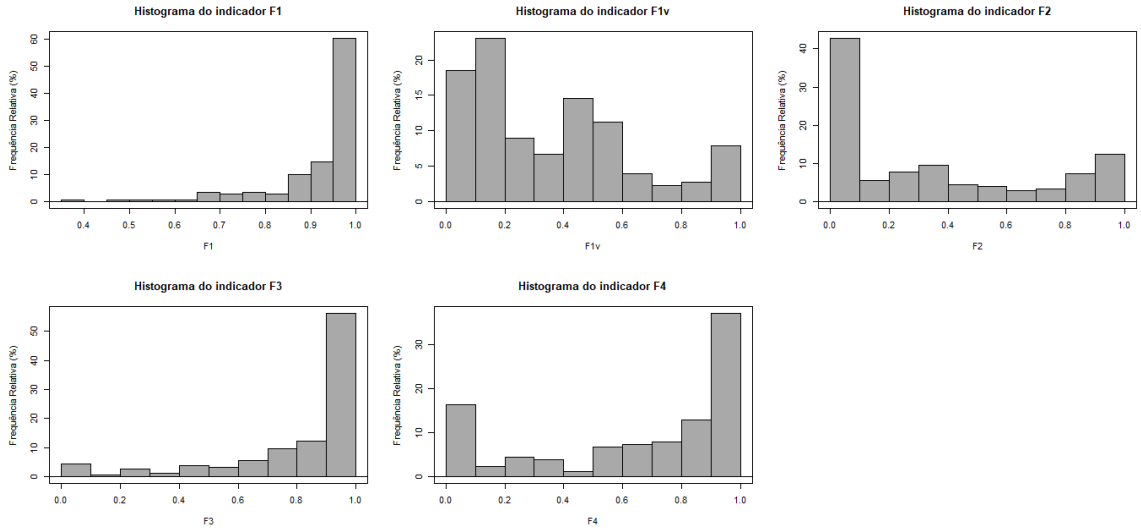

Figura 12 – Histograma dos indicadores F1, F1v, F2, F3 e F4.

No indicador F1, taxa máxima discriminante de Fisher, valores próximos a 1 indicam pouca influência do atributo mais discriminante na separação de classes, indicando maior complexidade do conjunto de dados. Medida complementar à de F1, a taxa máxima discriminante vetor-direcional de Fisher, ou F1v, busca encontrar um vetor que separe as duas classes, após os pontos do conjunto serem projetados nesse vetor. Valores maiores de F1v indicam menor complexidade. O indicador F2, volume da região de sobreposição, computa a sobreposição da distribuição dos valores dos atributos de cada classe do conjunto de dados. Valores maiores para a medida F2 indicam maior complexidade. Já o indicador F3, eficiência máxima individual do atributo, mede a eficiência de cada atributo em separar as classes. Menores valores de F3 indicam maior complexidade. Finalmente, o indicador F4, eficiência coletiva do atributo, indica a taxa de objetos discriminados pela combinação dos atributos do conjunto de dados. Valores maiores de F4 indicam maior eficiência discriminatória dos atributos.

Embora algumas das medidas baseadas em atributo meçam a complexidade numa relação inversa, na implementação do pacote ECoL esses indicadores tiveram seus valores invertidos, a saber, os indicadores F1v, F3 e F4. Algumas medidas de resumo dos valores

dos indicadores baseados em atributos para as bases de dados analisadas no experimento são:

```
F1                F1v               F2                F3                F4
Min.   :0.3729    Min.   :0.009753  Min.   :0.0000    Min.   :0.0000    Min.   :0.0000
1st Qu.:0.9003    1st Qu.:0.123940  1st Qu.:0.0000    1st Qu.:0.7315    1st Qu.:0.3804
Median :0.9602    Median :0.296264  Median :0.2095    Median :0.9234    Median :0.8018
Mean   :0.9190    Mean   :0.355141  Mean   :0.3313    Mean   :0.8103    Mean   :0.6465
3rd Qu.:0.9817    3rd Qu.:0.511639  3rd Qu.:0.6225    3rd Qu.:0.9819    3rd Qu.:0.9384
Max.   :1.0000    Max.   :0.998413  Max.   :1.0000    Max.   :1.0000    Max.   :1.0000
```

As medidas de resumo indicam que 75% dos conjuntos de dados possuem uma taxa F1 acima de 0,9, o que aponta para uma complexidade alta na separabilidade das classes desses conjuntos de dados.

```
       F1        F1v       F2        F3        F4
F1  1.0000000 0.5183323 0.2159469 0.6145837 0.4905165
F1v 0.5183323 1.0000000 0.3201774 0.4673509 0.5391104
F2  0.2159469 0.3201774 1.0000000 0.4854168 0.6004292
F3  0.6145837 0.4673509 0.4854168 1.0000000 0.8176911
F4  0.4905165 0.5391104 0.6004292 0.8176911 1.0000000
```

Analisando as relação entre os indicadores F1, F1v, F2, F3 e F4 nota-se uma correlação alta entre os indicadores F3 e F4.

Medidas de linearidade

Essa categoria de medidas de complexidade visa à quantificação da separabilidade das classes por um hiperplano, assumindo-se que um conjunto de dados cujas classes sejam linearmente separáveis é menos complexo que outro que exija um hiperplano não-linear para separar suas classes.

O indicador L1, soma do erro de distância por programação linear, calcula a média das distâncias dos objetos erroneamente classificados em relação ao hiperplano linear utilizado na classificação. Nesta medida, quanto maior o valor, mais complexo o problema, e os valores já são normalizados pelo pacote ECoL (GARCIA et al., 2020). O indicador L2, taxa de erro de classificador linear, calcula a taxa média de erro de um classificador SVM (*Support Vector Machines*), e apresenta valores maiores para problemas mais complexos. Por fim, o indicador L3, não-linearidade de um classificador linear, calcula a taxa de erro entre os pontos classificados por um classificador SVM (*Support Vector Machines*) treinado em conjunto de dados formado por pontos interpolados da mesma classe. Taxas mais altas do indicador L3 ocorrem em problemas mais complexos (LORENA et al., 2019, p.8).

A distribuição das frequências dos indicadores L1, L2 e L3 pode ser vista na Figura 13.

Algumas medidas de resumo dos valores dos indicadores de medidas de linearidade para as bases de dados analisadas no experimento são:

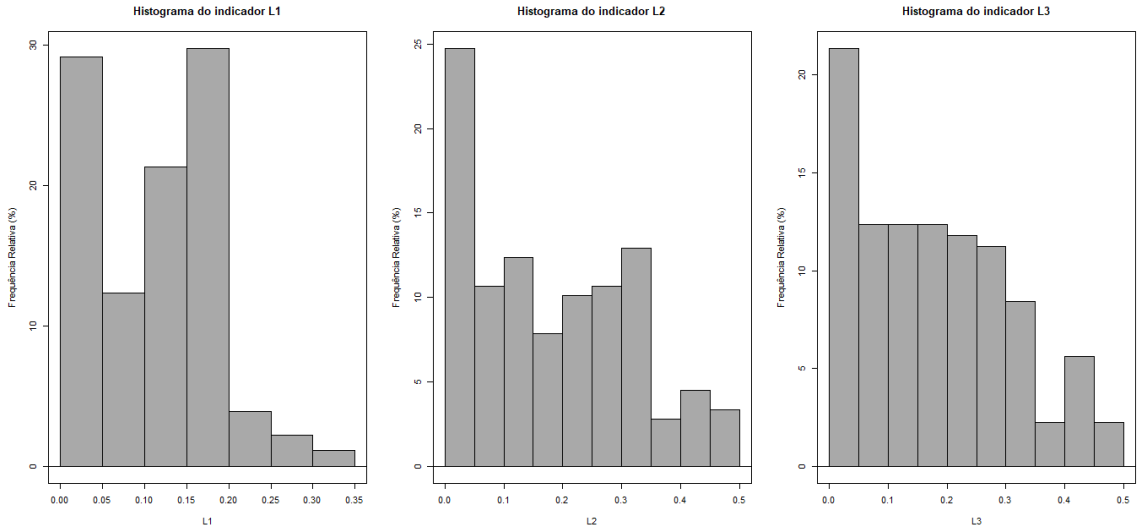

Figura 13 – Histograma dos indicadores L1, L2 e L3.

L1		L2		L3	
Min.	:0.00000	Min.	:0.00000	Min.	:0.00000
1st Qu.	:0.03509	1st Qu.	:0.05058	1st Qu.	:0.07287
Median	:0.11487	Median	:0.16533	Median	:0.15971
Mean	:0.11050	Mean	:0.18122	Mean	:0.17756
3rd Qu.	:0.16909	3rd Qu.	:0.28911	3rd Qu.	:0.27525
Max.	:0.31176	Max.	:0.48932	Max.	:0.48310

A correlação entre as variáveis L1, L2 e L3 é a seguinte:

	L1	L2	L3
L1	1.0000000	0.9054870	0.8189903
L2	0.9054870	1.0000000	0.9143344
L3	0.8189903	0.9143344	1.0000000

Medidas de vizinhança

Os indicadores representados pelas medidas de vizinhança buscam representar a forma da região de decisão e caracterizar a sobreposição dos pontos calculando a vizinhança local dos pontos (LORENA et al., 2019, pp.9-13).

O indicador N1, fração de pontos na fronteira da classe, indica o percentual de pontos que estão na borda ou em regiões de sobreposição das classes. Assim, quanto maior o valor de N1, mais complexo o conjunto de dados. O indicador N2, taxa média normalizada de distância NN extra/intra classe, calcula a taxa de distância intra-classe pela distância extra-classe dos pontos pela computação do vizinho mais próximo (NN). Valores maiores de N2 indicam problemas mais complexos. O indicador N3 calcula a taxa de erro média do classificador k-NN usando o procedimento *leave-one-out*. Valores mais altos indicam problemas mais complexos. No indicador N4, não-linearidade do classificador 1-NN, a taxa de erros é calculada aplicando-se o algoritmo k-NN em um conjunto de dados auxiliar composto por pontos interpolados do conjunto de dados original. Valores mais altos de

N4 ocorrem em conjuntos de dados mais complexos. O indicador N5 computa a fração de hiperesferas que agrupam dados de mesma classe em relação à quantidade de objetos do conjunto de dados, de modo que quanto mais hiperesferas forem necessárias, mais sobreposição de dados há e mais complexo é conjunto de dados. Finalmente, o indicador N6, cardinalidade média do conjunto local, computa a cardinalidade média de conjuntos de dados formados por objetos da mesma classe que o ponto analisado, dentro de um raio menor que a distância até o ponto mais próximo de outra classe. Embora valores menores para N6 sejam esperados para problemas mais complexos, o pacote ECoL já traz o resultado complementar (GARCIA et al., 2020).

A distribuição das frequências dos indicadores N1 a N6 pode ser vista na Figura 14.

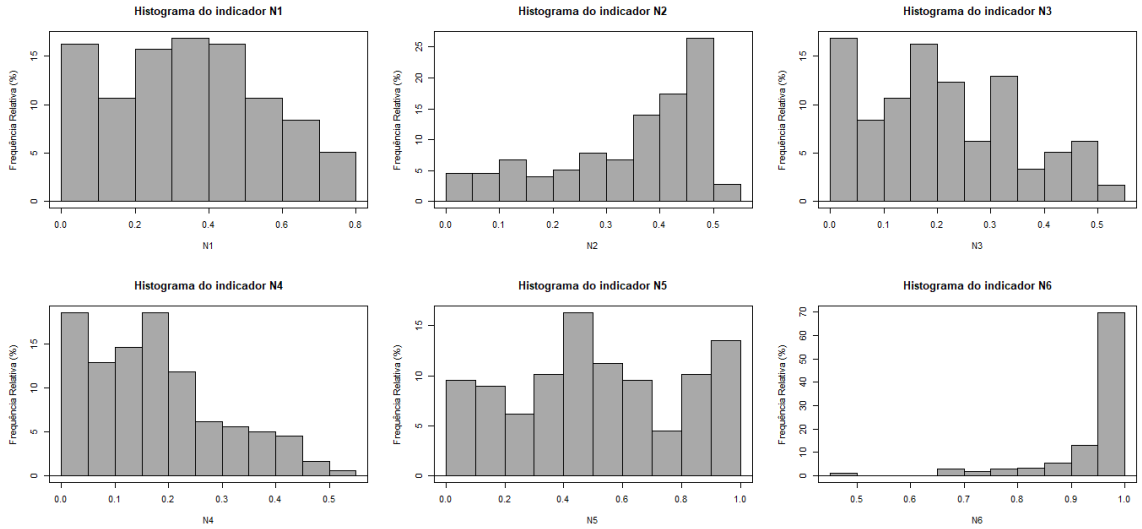

Figura 14 – Histograma dos indicadores N1 a N6.

Algumas medidas de resumo dos valores dos indicadores baseados em medidas de vizinhança para as bases de dados analisadas no experimento são:

```
N1                  N2                  N3
Min.   :0.003578    Min.   :0.0000    Min.   :0.00000
1st Qu.:0.191337    1st Qu.:0.2518    1st Qu.:0.09598
Median :0.349242    Median :0.3895    Median :0.19464
Mean   :0.347563    Mean   :0.3410    Mean   :0.21000
3rd Qu.:0.499194    3rd Qu.:0.4576    3rd Qu.:0.31547
Max.   :0.777580    Max.   :0.5203    Max.   :0.53500

N4                  N5                  N6
Min.   :0.00000     Min.   :0.0070    Min.   :0.4733
1st Qu.:0.07532     1st Qu.:0.3093    1st Qu.:0.9375
Median :0.15945     Median :0.5000    Median :0.9741
Mean   :0.17379     Mean   :0.5149    Mean   :0.9398
3rd Qu.:0.24017     3rd Qu.:0.7780    3rd Qu.:0.9899
Max.   :0.51335     Max.   :1.0000    Max.   :0.9999
```

Nota-se pelos valores de N6 que 75% dos conjuntos de dados analisados são considerados complexos, pela cálculo da cardinalidade média de conjuntos de dados locais.

A correlação entre os indicadores N1 a N6 calculou-se como:

	N1	N2	N3	N4	N5	N6
N1	1.0000000	0.6612767	0.8887456	0.6764132	0.7700221	0.6008595
N2	0.6612767	1.0000000	0.7738337	0.3258172	0.7848595	0.5970311
N3	0.8887456	0.7738337	1.0000000	0.6426352	0.7855695	0.5668171
N4	0.6764132	0.3258172	0.6426352	1.0000000	0.4296413	0.5402266
N5	0.7700221	0.7848595	0.7855695	0.4296413	1.0000000	0.6231703
N6	0.6008595	0.5970311	0.5668171	0.5402266	0.6231703	1.0000000

Medidas de rede

Nessa categoria de medidas, o conjunto de dados é representado como um grafo que preserva a distância entre os objetos. As arestas entre os vértices de diferentes classes são então podadas. A medida G1, densidade média da rede, calcula o número médio de arestas do grafo, obtendo valores mais altos para grafos mais densos, indicando problemas menos complexos. O pacote ECoL já retorna o valor complementar, indicando valores mais altos para conjuntos de dados mais complexos. A medida G2, coeficiente de agrupamento, calcula a taxa média do número de arestas entre vértices vizinhos pela quantidade máxima de arestas que poderiam existir entre os vizinhos, medindo a tendência dos vértices em se agruparem. Conjuntos de dados mais complexos retornarão valores menores para esse indicador, embora o pacote ECoL já retorne o valor complementar dessa medida. Finalmente, a medida G3, índice de pontos centrais, calcula a média dos índices de conexões de cada vértice do grafo, retornando valores menores para conjuntos de dados mais complexos. O pacote ECoL já retorna o valor complementar desse indicador (GARCIA et al., 2020; LORENA et al., 2019, pp.14,15).

Algumas medidas de resumo dos valores dos indicadores baseados em medidas de rede para as bases de dados analisadas no experimento são:

G1	G2	G3
Min. :0.8096	Min. :0.1189	Min. :0.2817
1st Qu.:0.8571	1st Qu.:0.3541	1st Qu.:0.6925
Median :0.8782	Median :0.4658	Median :0.7623
Mean :0.8740	Mean :0.4530	Mean :0.7462
3rd Qu.:0.8939	3rd Qu.:0.5546	3rd Qu.:0.8250
Max. :0.9649	Max. :1.0000	Max. :0.9431

A distribuição das frequências dos indicadores G1, G2 e G3 pode ser vista na Figura 15.

A correlação entre os indicadores foi calculada como:

	G1	G2	G3
G1	1.0000000	0.3670017	0.1684307
G2	0.3670017	1.0000000	−0.3991693
G3	0.1684307	−0.3991693	1.0000000

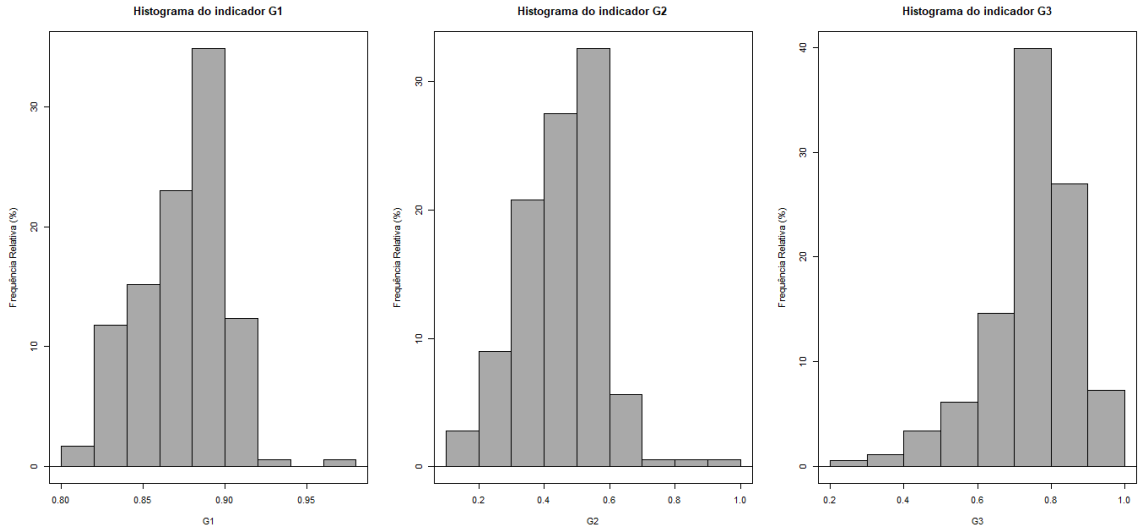

Figura 15 – Histograma dos indicadores G1, G2 e G3.

A correlação negativa entre os indicadores G3 e G2, embora seja média, não pôde ser explicada pela definição desses indicadores.

5.1.5 Atributos de qualidade da classificação

Como indicadores do construto Qualidade da Classificação foram utilizadas as medidas de AUC dos classificadores apresentados na Tabela 6. Os algoritmos de classificação selecionados são sensíveis às ocorrências de valores ausentes e discrepantes, capturando as variações dessas anomalias nos conjuntos de dados analisados.

As medidas de resumo para os indicadores de qualidade de classificação são:

C4.5		RF		CART	
Min.	:0.4833	Min.	:0.3605	Min.	:0.4420
1st Qu.	:0.6582	1st Qu.	:0.6778	1st Qu.	:0.6508
Median	:0.8033	Median	:0.8410	Median	:0.8084
Mean	:0.7724	Mean	:0.7969	Mean	:0.7714
3rd Qu.	:0.9069	3rd Qu.	:0.9296	3rd Qu.	:0.8677
Max.	:1.0000	Max.	:1.0000	Max.	:1.0000

Pela observação dos resumos percebe-se um bom desempenho desses classificadores em 50% dos conjuntos de dados analisados.

A distribuição das frequências dos indicadores dos classificadores C4.5, RF e CART pode ser vista na Figura 16.

A correlação entre os indicadores foi calculada como:

	C4.5	RF	CART
C4.5	1.0000000	0.8626815	0.8608670
RF	0.8626815	1.0000000	0.9287369
CART	0.8608670	0.9287369	1.0000000

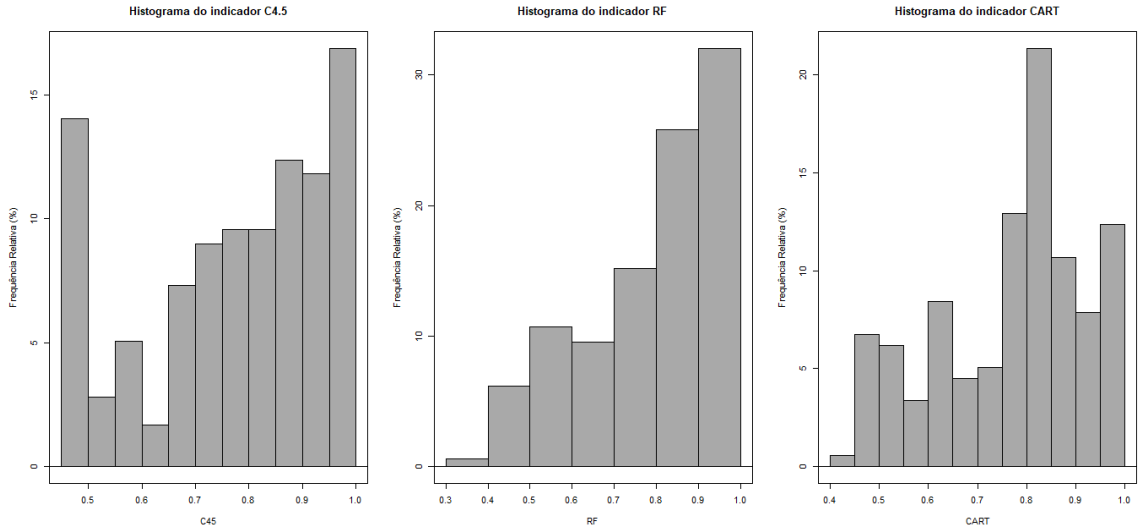

Figura 16 – Histograma dos indicadores dos classificadores C4.5, RF e CART.

Para os indicadores do construto Qualidade da Classificação, o valor do coeficiente alfa de Cronbach bem como o ganho de confiabilidade em caso de exclusão de algum dos indicadores, são apresentados a seguir:

```
Alpha reliability =  0.9571
Standardized alpha =  0.9581

Reliability deleting each item in turn:
      Alpha Std.Alpha r(item, total)
C4.5  0.962     0.963           0.878
CART  0.926     0.926           0.927
RF    0.923     0.925           0.927
```

Um valor de alfa de Cronbach de 0,958 é considerado um valor alto de confiabilidade para o modelo de mensuração do construto Qualidade da Classificação.

5.2 Avaliação dos resultados

A avaliação dos resultados do algoritmo PLS-SEM começa pela validação da qualidade dos modelos de mensuração reflexivos e formativos, e somente se as características de mensuração dos construtos forem aceitáveis é que os resultados do modelo estrutural poderão ser validados (HAIR et al., 2016, p.101).

5.2.1 Estimativa do modelo

Para o cálculo do modelo estrutural adotou-se a ferramenta SmartPLS 3.3.2 (RINGLE; WENDE; BECKER, 2015), configurada com os parâmetros abaixo, de acordo com Hair et al. (2016, pp.80-82):

- Atribuição de pesos: *path scheme*
- Padronização automática dos valores dos indicadores por escore-z
- Valor de inicialização dos relacionamentos do modelo de mensuração: +1
- Critério de parada/convergência do algoritmo: 10^{-5}
- Número máximo de iterações: 300

A estimativa inicial dos pesos, cargas, coeficientes e da variância foi efetuada utilizando-se os parâmetros mencionados anteriormente.

O modelo com os valores estimados é apresentado na Figura 17.

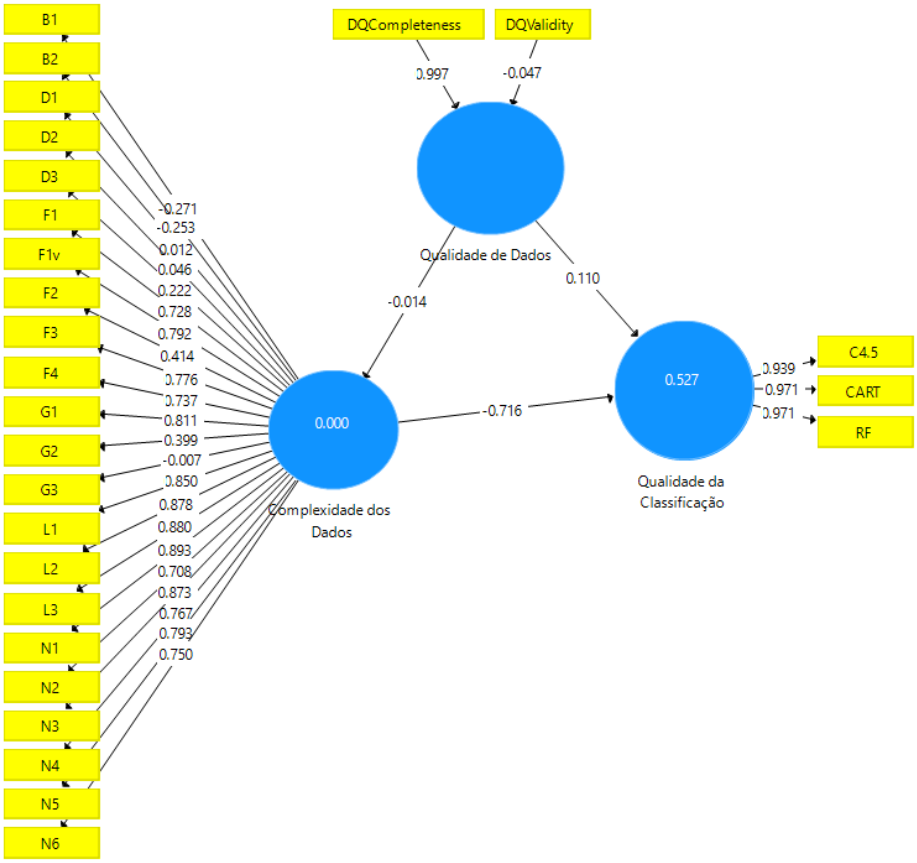

Figura 17 – Estimativa inicial do modelo.

5.2.2 Validação do modelo reflexivo

A avaliação do modelo reflexivo inicia-se pela avaliação da confiabilidade da consistência interna dos construtos. Para essa avaliação utilizam-se a medida alfa de Cronbach α (Equação 3.4) e o índice de confiabilidade composta ρ_c (Equação 3.5), cujos resultados estão apresentados na Tabela 9.

Tabela 9 – Confiabilidade dos construtos reflexivos.

	Alfa de Cronbach α	Confiabilidade composta ρ_c
Complexidade de Dados	0.889	0.918
Qualidade da Classificação	0.958	0.973

Embora valores altos para a confiabilidade composta ρ_c indiquem altos graus de confiabilidade, valores satisfatórios estão entre 0,70 e 0,90. Valores acima de 0,90 não são desejáveis, uma vez que ocorrem quando os indicadores do construto são redundantes (HAIR et al., 2016, p.102).

O pesquisador optou por **manter o construto Complexidade de Dados sem alterações**, esperando encontrar nas análises seguintes orientações mais claras sobre quais indicadores podem ser eliminados para diminuir a redundância das medidas. Quanto ao construto Qualidade da Classificação, optou-se por **manter os três indicadores da Tabela 6**, considerando que:

- simulações de retirada de indicadores não apresentaram ganho significativo nos valores de confiabilidade composta e alfa de Cronbach;
- embora os três algoritmos adotados, CART, C4.5 e *Random Forests*, sejam baseados em árvores e seus desempenhos de classificação sejam próximos, a permanência de suas medidas na análise é justificada pelo fato de se tratarem de algoritmos diferentes e pelo fato de não serem encontrados outros algoritmos de classificação que sejam ao mesmo tempo sensíveis a valores discrepantes e a valores ausentes.

A etapa seguinte de análise dos construtos reflexivos consiste em medir a validade convergente dos indicadores de cada construto, isto é, uma medida de quanto o indicador se correlaciona positivamente com outros indicadores do mesmo construto. Para essa avaliação são consideradas as cargas dos indicadores e também a Variância Média Extraída (AVE), calculada como na Equação 3.6. Embora para as cargas dos indicadores sejam esperados valores padronizadas de 0,708 ou maiores, Hair et al. (2016, pp.103,104) recomendam considerar o impacto da exclusão de um indicador com carga entre 0,40 e 0,708 na Variância Média Extraída e na confiabilidade composta.

Os valores para as cargas dos indicadores reflexivos dos construtos Complexidade dos Dados e Qualidade da Classificação são apresentados na Tabela 10.

Nota-se que os indicadores do construto Qualidade da Classificação apresentaram alta comunalidade, **sendo mantidos sem alteração**. No entanto, alguns indicadores do construto Complexidade dos Dados apresentaram cargas inferiores a 0,40 (linhas em negrito na Tabela 10). A pesquisa optou por **excluir os indicadores D1, D2, D3, G2 e G3**, avaliando a cada exclusão o ganho na confiabilidade composta e na Variância Média

Tabela 10 – Cargas dos indicadores reflexivos dos construtos Complexidade dos Dados e Qualidade da Classificação.

Indicador	**Carga**
Complexidade dos Dados	
B1	-0.271
B2	-0.253
D1	**0.012**
D2	**0.046**
D3	**0.222**
F1	0.728
F1v	0.792
F2	0.414
F3	0.776
F4	0.737
G1	0.811
G2	**0.399**
G3	-0.007
L1	0.850
L2	0.878
L3	0.880
N1	0.893
N2	0.708
N3	0.873
N4	0.767
N5	0.793
N6	0.750
Qualidade da Classificação	
RF	0.971
C4.5	0.939
CART	0.971

Extraída. Após essa exclusão, permaneceram negativas as cargas dos indicadores B1 e B2 (indicadores de desbalanceamento de classes) do construto Complexidade de Dados, valores não esperados, se consideradas as definições apresentadas na Subseção 3.2.6. A opção da pesquisa foi pela **exclusão dos indicadores B1 e B2**.

Os resultados resultantes para as confiabilidade e validade são apresentados na Tabela 11.

Tabela 11 – Confiabilidade e validade dos construtos reflexivos após a exclusão dos indicadores D1, D2, D3, G2, G3, B1 e B2.

	Alfa de Cronbach α	Confiabilidade composta ρ_c	Variância Média Extraída
Complexidade de Dados	0.956	0.961	0.625
Qualidade da Classificação	0.958	0.973	0.923

Os valores para as cargas dos indicadores reflexivos dos construtos Complexidade dos Dados e Qualidade da Classificação após as exclusões dos indicadores D1, D2, D3, G2, G3, B1 e B2 são apresentados na Tabela 12.

Embora altos, os valores de confiabilidade e validade foram considerados aceitáveis

Tabela 12 – Cargas dos indicadores reflexivos dos construtos Complexidade dos Dados e Qualidade da Classificação após exclusões dos indicadores D1, D2, D3, G2, G3, B1 e B2.

Indicador	**Carga**
Complexidade dos Dados	
F1	0.699
F1v	0.810
F2	0.438
F3	0.770
F4	0.755
G1	0.817
L1	0.875
L2	0.905
L3	0.896
N1	0.887
N2	0.702
N3	0.873
N4	0.764
N5	0.783
N6	0.762
Qualidade da Classificação	
RF	0.971
C4.5	0.939
CART	0.972

para as análises, considerando-se que:

- a pesquisa dispõe de apenas três indicadores para medir a qualidade da classificação, representando os poucos algoritmos de classificação sensíveis a valores discrepantes e a valores ausentes;
- os diferentes aspectos da complexidade dos dados discutidos por Lorena et al. (2019) foram resumidos em apenas uma variável latente;
- os valores de Variância Média Extraída para os construtos são acima de 0,5, sendo considerados aceitáveis (HAIR et al., 2016, p.107)

Por fim, a análise dos construtos reflexivos passa pela avaliação da sua validade discriminante, isto é, do quanto o construto é distinto de outros construtos. Para essa validação são analisadas as cargas cruzadas dos indicadores, que devem ser maiores que todas suas cargas em outros construtos, e os resultados do critério Fornell-Larcker (HAIR et al., 2016, p.105).

Os valores para as cargas cruzadas dos indicadores são apresentados na Tabela 13.

Os valores da análise discriminante pelo critério de Fornell-Larcker são apresentados na Tabela 14.

Conclui-se pelos resultados das cargas cruzadas e pelos valores da análise discriminante pelo critério de Fornell-Larcker que os construtos Qualidade da Classificação e Complexidade dos Dados são distintos entre si, isto é, medem fenômenos diferentes.

Tabela 13 – Cargas cruzadas (destacadas em negrito) dos indicadores reflexivos dos construtos Complexidade dos Dados e Qualidade da Classificação.

Indicador	**Complexidade dos Dados**	**Qualidade da Classificação**
F1	**0.699**	-0.505
F1v	**0.810**	-0.648
F2	**0.438**	-0.054
F3	**0.770**	-0.499
F4	**0.755**	-0.407
G1	**0.817**	-0.431
L1	**0.875**	-0.566
L2	**0.905**	-0.574
L3	**0.896**	-0.565
N1	**0.887**	-0.692
N2	**0.702**	-0.477
N3	**0.873**	-0.737
N4	**0.764**	-0.599
N5	**0.783**	-0.470
N6	**0.762**	-0.373
RF	-0.672	**0.971**
C4.5	-0.575	**0.939**
CART	-0.708	**0.972**

Tabela 14 – Valores da análise discriminante pelo critério de Fornell-Larcker para os construtos reflexivos.

	Complexidade dos Dados	**Qualidade da Classificação**
Complexidade dos Dados	**0.791**	
Qualidade da Classificação	-0.682	**0.961**
Qualidade dos Dados	0.051	0.118

5.2.3 Validação do modelo formativo

A validação do modelo formativo seguiu os passos descritos na Subseção 3.1.3.2, avaliando aspectos diferentes dos aspectos avaliados nos modelos reflexivos.

O primeiro instrumento para a validação do modelo formativo é a verificação da validade convergente, que é a extensão pela qual o construto formativo se correlaciona com outro construto reflexivo de um único item que capture o mesmo conceito e que utilize diferentes indicadores. O construto reflexivo de um único item global que capture o mesmo conceito é definido na fase de projeto nas pesquisas em Ciências Sociais, mas **não existe na presente pesquisa pelo fato de que todos os atributos de qualidade de dados de interesse da pesquisa formam o construto Qualidade de Dados**.

O segundo instrumento utilizado para validar o modelo formativo é pela colinearidade dos seus indicadores, isto é, uma alta correlação entre eles. Uma vez que indicadores formativos medem aspectos diferentes do fenômeno medido pelo construto, não é esperado haver uma alta correlação entre eles. O indicador VIF (*variance inflation factor*), calculado como na Equação 3.8, mede o grau para o qual o erro padrão aumenta devido a presença de colinearidade. Para valores de colinearidade considerados não críticos (VIF

< 5), a significância dos pesos e a contribuição do indicador devem ser analisados (HAIR et al., 2016, pp.125,126).

A validação do modelo formativo passa pela avaliação da contribuição do indicador, expressa por seu peso. Além de serem comparados entre si para calcular sua contribuição relativa, os pesos dos indicadores são testados pela abordagem de *bootstrapping* para verificar se são significativamente diferentes de zero. Para a execução do *bootstrapping* foram geradas 5.000 sub-amostras ($\alpha = 0.05$, teste bicaudal).

Os histogramas dos coeficientes do modelo estrutural interno são unimodais para os três relacionamentos entre construtos, como ilustram as Figuras 18, 19 e 20.

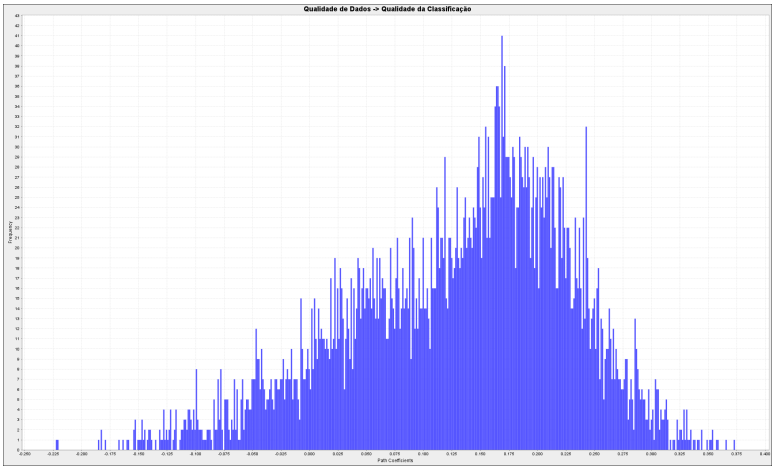

Figura 18 – Histograma dos coeficientes do relacionamento entre os construtos Qualidade de Dados e Qualidade da Classificação obtidos no processo de *bootstrapping*.

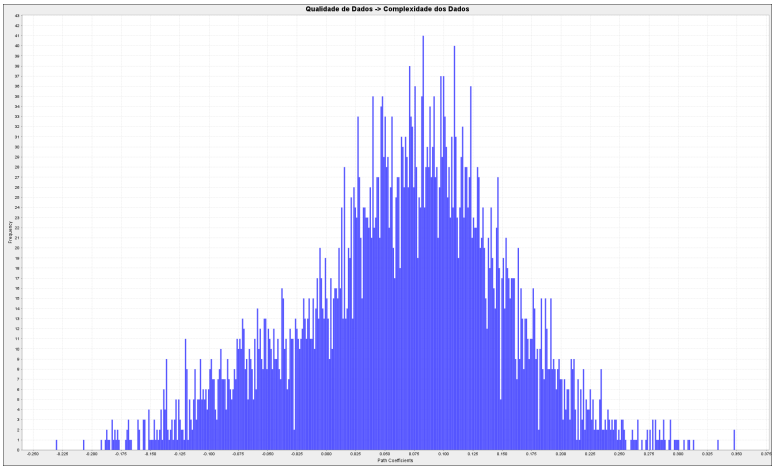

Figura 19 – Histograma dos coeficientes do relacionamento entre os construtos Qualidade de Dados e Complexidade de Dados obtidos no processo de *bootstrapping*.

Os valores de colinearidade, dos pesos originais e da significância após o *bootstrapping* encontrados para os indicadores formativos do construto Qualidade de Dados estão representados na Tabela 15.

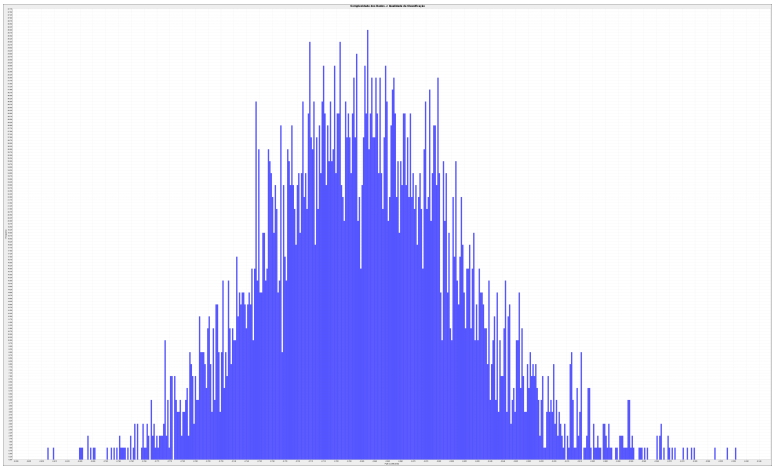

Figura 20 – Histograma dos coeficientes do relacionamento entre os construtos Complexidade de Dados e Qualidade da Classificação obtidos no processo de *bootstrapping*.

Tabela 15 – Valores de colinearidade, dos pesos e da significância para os indicadores do construto Qualidade de Dados.

Indicador	**VIF**	**Peso**	***valor-t***	**Intervalo de confiança**
DQCompleteness	1.002	0.993	1.999	[0.0189;1.9671]
DQValidity	1.002	0.162	0.348	[-0.7494;1.0734]

Para o indicador DQCompleteness, o peso de 0,993 foi considerado significantemente diferente de zero, a um grau de significância de 0,05, para o teste bicaudal (*two tailed*). Já para o indicador DQValidity o peso de 0,162 foi considerado não significativo, a um grau de significância 0,05, para o teste bicaudal. No entanto, **optou-se pela manutenção do indicador DQValidity para manter o domínio do construto compatível com a teoria**.

5.2.4 Validação do modelo estrutural

Nessa etapa validou-se o modelo estrutural que representa a relação entre qualidade de dados, complexidade de dados e qualidade da classificação. Os resultados permitem entender quão bem os dados empíricos suportam os conceitos propostos no modelo estrutural e verificar se esses conceitos foram empiricamente confirmados.

O passo inicial de análise é a busca por colinearidade no modelo estrutural, que é identificada com valores de VIF acima de 5. O passo seguinte visa à validação da significância e da relevância dos relacionamentos do modelo estrutural. Segue-se, então, a avaliação da variância explicada (R^2) e do tamanho do efeito (f^2). Essas medidas avaliam a capacidade preditiva do modelo estrutural (HAIR et al., 2016, p.169).

A Tabela 16 apresenta os os resultados e indicadores de significância dos relacionamentos propostos pelo modelo estrutural.

Os valores apresentados na Tabela 16 indicam que não foi encontrada colinearidade

Tabela 16 – Resultados e indicadores de significância dos relacionamentos propostos pelo modelo estrutural.

Relacionamento	**VIF**	f^2	**Coeficiente estrutural padronizado**	**Desvio-padrão**	***valor-t***	***valor-p***	Q^2	R^2 **ajustado**
Complexidade dos Dados -> Qualidade da Classificação	1.003	0,930	-0,690	0,038	18,08	0,000	0,438	0,483
Qualidade de Dados -> Qualidade da Classificação	1,003	0,046	0,154	0,097	1,59	0,113		
Qualidade de Dados -> Complexidade dos Dados	1,000	0,003	0,051	0,086	0,60	0,552	0,001	0,003

nos conjuntos de variáveis preditoras. Os valores dos coeficientes estruturais padronizados representam os coeficientes dos relacionamentos entre os construtos, com um valor padronizado variando entre -1 e 1. Os valores do coeficiente estrutural padronizado encontrados para os relacionamentos do construto Qualidade de Dados com o construto Qualidade da Classificação (0,154) e também com o construto Complexidade dos Dados (0,051), indicam **relacionamentos fracos**, que também **não são significantes**, uma vez que seus valores-t são de 1,59 e 0,60, respectivamente, para um grau de significância de 5%. Já o relacionamento entre os construtos Complexidade dos Dados e Qualidade da Classificação tem um coeficiente padronizado de -0,690 com um valor-t de 18,08, para um grau de significância de 5%.

Os valores do R^2 ajustado representam os efeitos combinados dos construtos exógenos sobre os construtos endógenos, e representam também a quantidade de variância nos construtos endógenos explicada pelos construtos endógenos conectados a eles. Os resultados apresentados na Tabela 16 informam que o construto Qualidade de Dados explicou apenas 0,3% da variância do construto Complexidade de Dados, ao passo que os construtos Qualidade de Dados e Complexidade de Dados explicaram 48,3 % da variância do construto Qualidade da Classificação. O valor de Q^2, obtido pelo processo de *blindfolding* com uma distância de omissão 7, confirma a relevância preditiva do modelo para prever os pontos do construto Qualidade da Classificação.

Na Figura 21 os caminhos entre os construtos estão destacados proporcionalmente à sua contribuição para o resultado do construto endógeno.

5.3 Construção do indicador de qualidade de classificação

A construção de um indicador de qualidade para classificações binárias tem como meta a abstração de aspectos multidimensionais de conjuntos de dados a serem submetidos ao processo de classificação, oferecendo ao analista de dados uma medida de fácil interpretação e acessível para comparações. Um indicador artificial, também conhecido

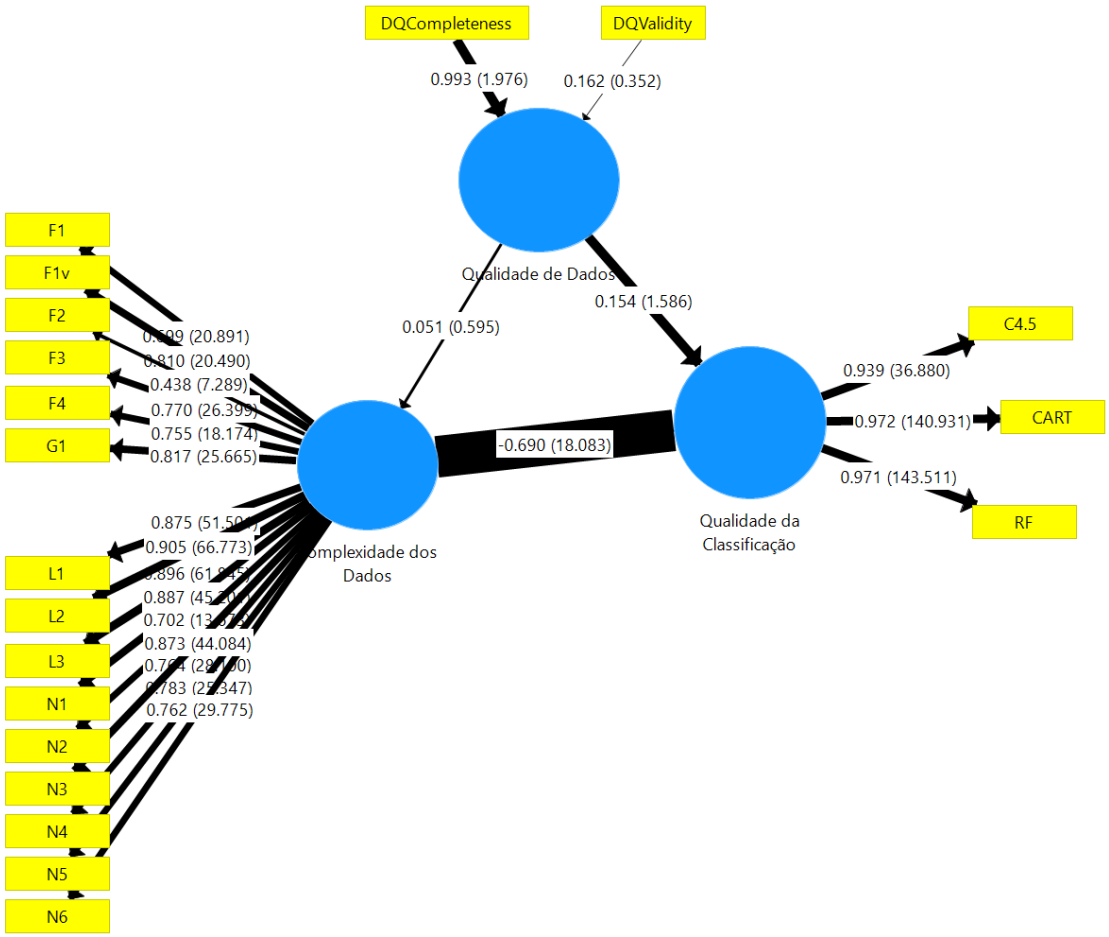

Figura 21 – Caminhos entre os construtos com destaque proporcional à sua contribuição no modelo.

como composto, ou ainda, sintético, pode ser entendido como uma medida qualitativa ou quantitativa derivada de dados observáveis e que pode apresentar posições relativas em uma determinada área do conhecimento (NARDO et al., 2008).

Embora o desenvolvimento de indicadores compostos seja uma prática em diversas áreas do conhecimento (GRECO et al., 2018), o benefício de concentrar a complexidade dos componentes individuais em uma medida simples pode ocultar falhas na sua construção e abrir oportunidades para interpretações equivocadas de seus resultados. Visando à diminuição do risco de falta de transparência em sua construção, Nardo et al. (2008) propõem recomendações sobre como projetar, desenvolver e disseminar indicadores compostos, apresentando os seguintes passos:

- desenvolvimento do quadro teórico que fundamentará a seleção e a combinação de indicadores que formarão um indicador composto significativo
- indicadores relevantes para o fenômeno de interesse devem ser medidos e relacionados uns aos outros, formando o conjunto de dados de análise
- dados discrepantes e ausentes devem ser tratados

- a análise exploratória dos dados deve ser aplicada para investigar a estrutura dos dados e ajudar a definir os métodos de ponderação e agregação dos indicadores
- o conjunto de dados deve ser normalizado para permitir comparação
- os processos de ponderação e de agregação dos indicadores devem ser executados de acordo com o quadro teórico que fundamentou a seleção dos indicadores
- análises de robustez e sensitividade do índice devem ser aplicadas pela variação de indicadores e dos métodos de imputação de valores ausentes, de normalização, de ponderação e de agregação
- a decomposição do indicador deve ser realizada para analisar a contribuição de seus sub-componentes para o resultado final
- tentativas de relacionar o indicador composto com outros indicadores similares devem ser feitas para identificar conexões entre os indicadores
- validações da representação visual do indicador devem ser feitas visando a uma correta interpretação do indicador

Parte dos passos apresentados por Nardo et al. (2008) para a proposição de um indicador composto se sobrepõe às atividades executadas no presente trabalho na proposta de um modelo estrutural, a saber: o desenvolvimento do quadro teórico, a seleção dos indicadores, a medição dos indicadores e a composição do conjunto de dados, o tratamento de dados ausentes e discrepantes, a análise exploratória do conjunto de dados, a normalização dos dados e a definição do método de ponderação dos indicadores. O uso da Modelagem de Equações Estruturais e do algoritmo PLS-SEM como ferramentas para o relacionamento dos indicadores e como método de ponderação dos indicadores individuais encontra precedente na literatura (NARDO et al., 2008; LIBóRIO et al., 2020; TOMASELLI; FORDELLONE; VICHI, 2020) e se mostra como uma ferramenta útil para se trabalhar com indicadores compostos pela possibilidade de se incluir variáveis latentes como fatores (NARDO et al., 2008, p.135) e pela capacidade de medir a força e a significância dos efeitos entre os indicadores individuais sobre o indicador composto (LIBóRIO et al., 2020).

5.3.1 Definição do conceito do indicador composto

O indicador composto proposto pela pesquisa, chamado a partir desse ponto de IQCb, ou Indicador da Qualidade da Classificação de conjuntos de dados binários, tem a sua fundamentação discutida no Capítulo 4.1.2. A ideia subjacente ao IQCb é que a qualidade da classificação de um conjunto de dados binário é dependente da qualidade

e da complexidade dos seus dados, ou, na forma de função, o indicador IQCb para o conjunto de dados $d, d = 1, ..., M$, pode ser representado como (Equação 5.1):

$$IQCb_d = f(I_{1,d}, I_{2,d}, ...I_{Q,d}, w_1, w_2, ...w_Q) \tag{5.1}$$

onde f representa a metodologia de agregação adotada (ver Seção 5.3.2) para os Q indicadores individuais $I_{1,d}, I_{2,d}, ...I_{Q,d}$ ponderados por seus respectivos pesos $w_1, w_2, ...w_Q$.

Numa visão hierárquica o IQCb poderia ser organizado como na Figura 22.

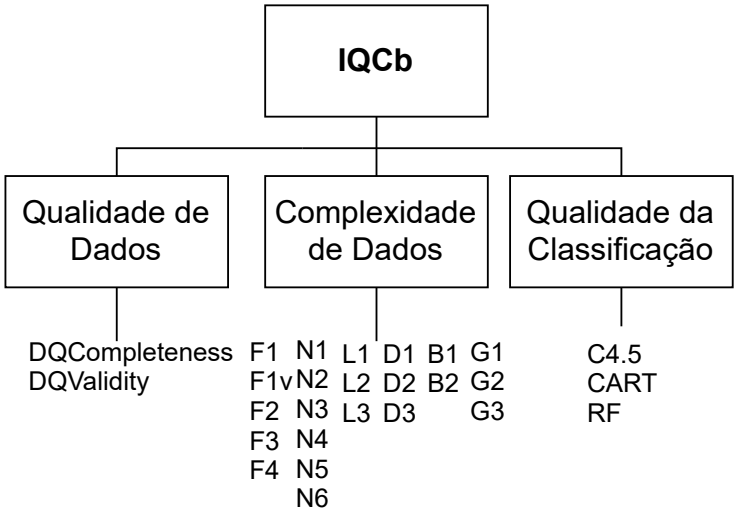

Figura 22 – Representação hierárquica do Indicador de Qualidade da Classificação (IQCb). No primeiro nível está o indicador composto, no segundo nível estão os indicadores individuais (construtos) e no terceiro nível, os indicadores que formam os indicadores individuais.

5.3.2 Metodologia de ponderação dos indicadores individuais

Os indicadores individuais, que na Modelagem de Equações Estruturais são comparáveis às variáveis latentes, contribuem com graus diferentes de importância para o resultado final do indicador composto. A definição da contribuição relativa de cada indicador individual, ou ponderação, foi calculada na presente pesquisa pela aplicação do algoritmo PLS-SEM. Outros métodos de ponderação poderiam ser aplicados na construção do indicador composto, resultando em pesos diferentes para os indicadores individuais (NARDO et al., 2008).

A atribuição dos pesos para os indicadores individuais pode ser calculada com base nos resultados dos coeficientes estruturais padronizados dos relacionamentos entre as variáveis latentes, apresentados na Tabela 16 e na Figura 21, e sintetizados na Tabela 17.

A opção do pesquisador foi a de considerar os coeficientes estruturais padronizados como significantes e seguir a abordagem de Libório et al. (2020, p.11) em calcular a

Tabela 17 – Coeficientes estruturais padronizados para os relacionamentos entre as variáveis latentes, reproduzidos da Tabela 16 e da Figura 21.

Relacionamento	**Coeficiente estrutural padronizado**
Complexidade dos Dados -> Qualidade da Classificação	-0,690
Qualidade de Dados -> Qualidade da Classificação	0,154
Qualidade de Dados -> Complexidade dos Dados	0,051

contribuição proporcional das variáveis latentes Complexidade de Dados e Qualidade de Dados para a Qualidade da Classificação. Nessa abordagem, a contribuição proporcional é calculada pela divisão do módulo do valor do coeficiente estrutural padronizado pela soma do módulo dos coeficientes, considerando apenas os coeficientes dos relacionamentos diretos com a variável latente Qualidade de Classificação. O módulo do valor do coeficiente estrutural padronizado é utilizado porque o sinal representa a influência de um construto em outro, que pode ser uma influência negativa (quanto maior o valor de um construto, menor o valor do outro) ou positiva (quanto maior o valor de umm construto, maior o valor do outro). Assim os pesos serão (Tabela 18):

Tabela 18 – Definição dos indicadores e cálculo dos pesos dos relacionamentos diretos entre variáveis latentes, calculados com base nos coeficientes estruturais padronizados.

Relacionamento	**Coeficiente estrutural padronizado (em módulo)**	**Indicador (I)**	**Peso proporcional (w)**
Complexidade dos Dados -> Qualidade da Classificação	0,690	DC	$0,690/(0,690+0,154) = 0,817$
Qualidade de Dados -> Qualidade da Classificação	0,154	DQ	$0,154/(0,690+0,154) = 0,183$

5.3.3 Metodologia de agregação

A metodologia de agregação dos indicadores individuais para a composição do indicador deve procurar representar adequadamente a relação entre os indicadores individuais, com base no quadro teórico. Para representar as influências conjuntas de qualidade e complexidade dos dados sobre a qualidade da classificação, conforme explicado no Capítulo 4.1, uma metodologia de agregação menos compensatória parece ser mais apropriada (NARDO et al., 2008, p.33).

Numa metodologia de agregação, o desempenho insatisfatório em uma dimensão pode ser compensado por valores suficientemente altos em outra dimensão, o que distorce a importância real de uma variável representada pelo peso a ela atribuída (NARDO et al., 2008, pp.33,103). A opção do pesquisador foi a de adotar uma metodologia de agregação de indicadores por média geométrica, de menor compensabilidade para indicadores com valores baixos, a exemplo do adotado no indicador CPI-U-XG (*Consumer Price Index for All Urban Consumers using geometric means*) (STATISTICS, 1999). Assim, o cálculo do

indicador composto IQCb para o conjunto de dados $d, d = 1, ..., M$ e para o metodologia de agregação por média geométrica pode ser representado como (Equação 5.2):

$$IQCb_d = (\prod_{q=1}^{Q} I_{i,d} w_i)^{\frac{1}{Q}} = \sqrt[Q]{I_{1,d} w_1 \cdot I_{2,d} w_2 \cdot ... \cdot I_{Q,d} w_Q} \tag{5.2}$$

para os Q indicadores individuais $I_{1,d}, I_{2,d}, ...I_{Q,d}$ ponderados por seus respectivos pesos $w_1, w_2, ...w_Q$.

5.3.4 Implementação do IQCb

Se os indicadores DC e DQ da Tabela 18 substituírem os indicadores $I_{q,d}$ na Equação 5.2 e valores $0,817$ e $0,183$ substituírem os pesos w_q, o cálculo do indicador composto IQCb para o conjunto de dados $d, d = 1, ..., M$, resultará na equação (Equação 5.3):

$$IQCb_d = \sqrt[2]{0.817 DC_d \cdot 0.183 DQ_d} \tag{5.3}$$

5.3.5 Aplicabilidade

Conhecer a proporção em que a qualidade da classificação é afetada pela qualidade e pela complexidade dos dados permite antecipar o resultado de uma análise de classificação. Antes de submeter um conjunto de dados para análise, é possível coletar os indicadores de Qualidade de Dados e Complexidade de Dados para mostrar a tendência do resultado da tarefa de classificação: valores altos para indicadores de Complexidade de Dados podem sugerir um desempenho insatisfatório para classificadores neste conjunto de dados.

A contribuição do modelo gerado pelo algoritmo PLS-SEM para o relacionamento entre Complexidade de Dados (DC), Qualidade de Dados (DQ) e Qualidade de Classificação (CQ) é permitir uma estimativa mais precisa da influência da qualidade e complexidade dos dados no resultado final da qualidade de classificação. Embora seja possível interpretar a expectativa de desempenho de análises observando os valores dos indicadores de Qualidade de Dados e de Complexidade de dados, o indicador composto como o indicador de qualidade da classificação, IQCb, permite uma interpretação direta dessa expectativa.

Na Figura 23 quatro conjuntos de dados do repositório OpenML (CASALICCHIO et al., 2017) foram submetidos aos procedimentos de coleta de metadados de qualidade e complexidade de dados descritos nas seções 3.3 e 3.2. As medidas dos metadados de qualidade e complexidade foram agrupadas e ponderadas para representar os indicadores de Qualidade de Dados (DQ) e Complexidade de Dados (DC), seguindo as cargas e os pesos representados na Figura 21. Os cálculos são demonstrados na Tabela 19.

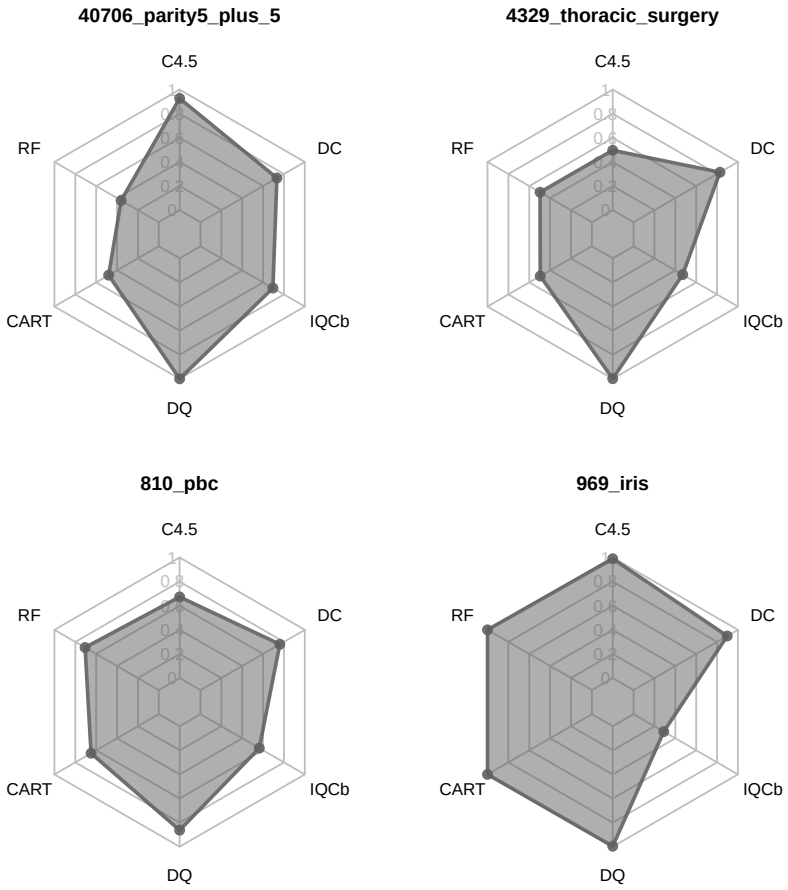

Figura 23 – Representação gráfica de resultados de algoritmos de classificação (RF, C4.5 e CART), do indicador de Qualidade de Dados (DQ), do indicador de Complexidade de Dados (DC) e do Indicador de Qualidade da Classificação (IQCb) para quatro conjuntos de dados do repositório OpenML.

Tabela 19 – Cálculo de medidas agrupadas de Complexidade e Qualidade de Dados, pelo critério de média ponderada. Os pesos foram obtidos a partir do modelo gerado pelo algoritmo PLS-SEM (Figura 21).

Indicador	**Cálculo (média ponderada dos metadados)**
DC (Complexidade de Dados)	$(F1*0.699)+(F1v*0.810)+(F2*0.438)+(F3*0.770)+(F4*0.755)/(0.699+0.810+0.438+0.770+0.755)+(N1*0.887)+(N2*0.702)+(N3*0.873)+(N4*0.764)+(N5*0.783)+(N6*0.762))/(0.887+0.702+0.873+0.764+0.783+0.762)+(L1*0.875)+(L2*0.905)+(L3*0.896))/(0.875+0.905+0.896)+(G1*0.755)/(0.755))$
DQ (Qualidade de Dados)	$(DQCompleteness*0.993)+(DQValidity*0.162))/(0.993+0.162)$
IQCb	$\sqrt[2]{0,817DC \cdot 0,183DQ}$

Os valores para os indicadores da Figura 23 para os quatro conjuntos de dados são apresentados na Tabela 20.

Tabela 20 – Valores para os indicadores da Figura 23.

Conjunto de Dados	**C4.5**	**RF**	**CART**	**DQ**	**DC**	**IQCb**
40706_parity5_plus_5	0.9269523	0.3604648	0.4803281	1.0000000	0.7324424	0.6939045
810_pbc	0.6730740	0.7060577	0.6485062	0.8625198	0.7602019	0.5635475
4329_thoracic_surgery	0.4962500	0.4962500	0.4946429	0.9991047	0.8278862	0.4705198
969_iris	0.9900000	1.0000000	1.0000000	0.9962597	0.8975554	0.2891505

Entre os conjuntos de dados selecionados como exemplo, os casos em que a complexidade de dados é menor os resultados dos classificadores C4.5, RF e CART foram melhores, exceto pelo conjunto de dados "4329_thoracic_surgery". De fato há uma correlação positiva entre o indicador composto IQCb e os resultados das classificações (Figura 24).

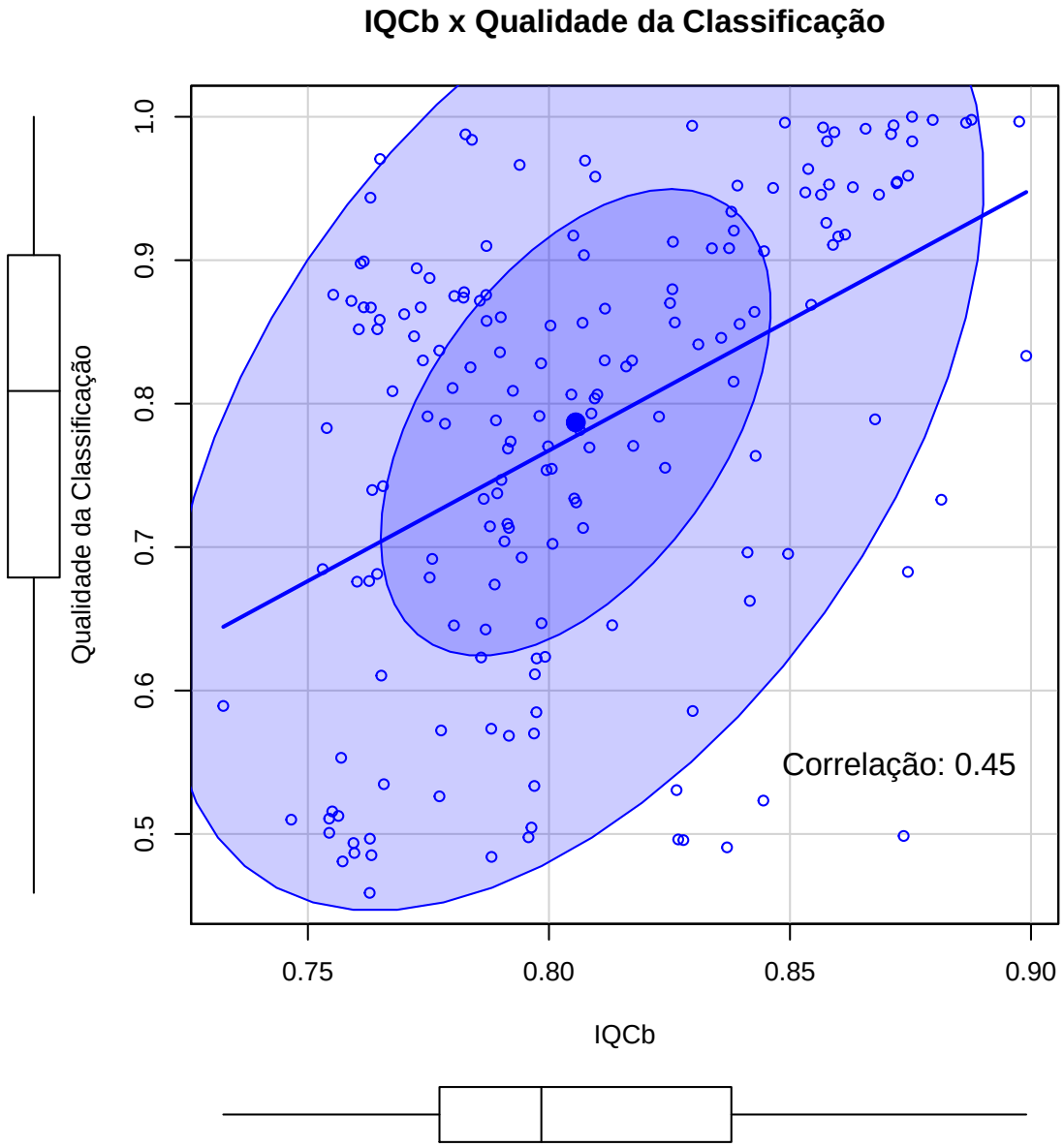

Figura 24 – Gráfico de dispersão entre o resultado médio da Qualidade da Classificação (RF, C4.5 e CART) e o indicador composto IQCb para os 178 conjuntos de dados do repositório OpenML analisados na pesquisa.

A Figura 24 apresenta a dispersão entre os resultados médios da variável independente Qualidade de Classificação, calculada pela média aritmética dos indicadores RF, C4.5 e CART, e os resultados do indicador composto IQCb, para os 178 conjuntos de dados do repositório OpenML analisados na pesquisa. As elipses indicam a concentração dos pontos com graus 0,5 e 0,95 graus de contorno de probabilidade normal, o ponto central

indica o centro das elipses. Ao lado dos eixos apresentam-se diagramas de caixa (*boxplots*). A correlação encontrada entre os resultados médios da variável independente Qualidade de Classificação e os resultados do indicador composto IQCb foi de **0,45**, indicando que o indicador composto IQCb conseguiu medir a variação da qualidade da classificação.

6 Conclusões, Limitações da Pesquisa e Trabalhos Futuros

A qualidade dos dados é uma preocupação real e bem representada no processo de Descoberta de Conhecimento em Bases de Dados. Um melhor entendimento da relação Qualidade de Dados e Complexidade de Dados pode trazer ganhos de qualidade às análises, o que não deve ser ignorado em uma realidade de Big Data.

A pesquisa se mostrou inovadora ao relacionar aspectos dos dados que, em geral, são tratados separadamente e cujo efeito é quase ignorado: A Qualidade dos Dados afeta a Complexidade dos Dados e ambos afetam a Qualidade da Classificação. Validade e completude se mostraram no modelo proposto como dois importantes problemas de qualidade cujo efeito sobre a Complexidade dos Dados merece ser estudado mais profundamente. Além disso, o uso da Modelagem de Equações Estruturais e do algoritmo PLS-SEM para estudar as relações entre as dimensões de qualidade e complexidade, até onde se sabe, é inédito na literatura, e abriu uma nova plataforma para este ferramental nas áreas de Mineração de Dados, Big Data e Governança de Dados.

O uso do PLS-SEM permitiu a quantificação da contribuição combinada da qualidade e complexidade dos dados para o sucesso das classificações nos conjuntos de dados de duas classes. Os resultados sugerem que os fatores estruturais que interferem na complexidade de um conjunto de dados merecem mais atenção nos problemas de classificação do que a ocorrência de valores ausentes e discrepâncias, e isso requer um maior investimento de tempo do analista nas etapas de pré-processamento.

A construção de um indicador composto que permite a previsão da qualidade dos resultados de classificação com base nos metadados de qualidade e complexidade dos conjuntos de dados a serem analisados se mostrou possível. O indicador IQCb pode ser aplicado como uma medida que permite a abstração da complexidade dos componentes do indicador a comparação de desempenho dos conjuntos de dados.

Oportunidades de continuação se apresentam em alguns pontos da pesquisa. A categorização de valores ausentes proposta por Rubin (1976) indica que uma medida simples de completude pode não ser tão adequada. A ocorrência aleatória de valores ausentes torna o conjunto de dados mais adequado às análises do que a ocorrência da mesma quantidade de valores ausentes em um pouco número de variáveis, ou em variáveis mais significativas para a análise (HAIR et al., 2014, p.45). Assim, parece que a completude de um conjunto de dados não é dependente apenas da quantidade de valores ausentes, mas também da distribuição das ausências. Dessa forma, a proposição de um modelo que

faça uso de medidas de variância das ausências e de proporção de valores ausentes pode eventualmente alcançar melhores resultados na construção de um indicador de qualidade das classificações.

Uma importante oportunidade de pesquisa seria ainda a construção artificial de um conjunto de dados experimental. Embora conjuntos de dados estruturados (reais) difiram de conjuntos de dados aleatórios na construção de classificadores que possam prever corretamente uma classe (HO; BASU, 2000) a utilização de um conjunto de dados artificialmente construído permitiria maior controle sobre variáveis às quais os algoritmos são mais sensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p.11), tais como dimensionalidade e esparsamento dos dados e permitiria, eventualmente, maior conhecimento do comportamento das variáveis latentes de qualidade e complexidade.

A pesquisa pode ser desenvolvida também pela inclusão de outros algoritmos não baseados em árvore, tais como classificadores baseados em vizinhança e classificadores lineares, mas que sejam sensíveis a valores discrepantes e a ausências como indicadores de medição do desempenho da classificação. Ainda, na construção do conjunto de dados experimental a adoção de uma estratégia de decomposição de problemas multiclasses em problemas binários, a exemplo da estratégia OVO (*One-vs-One*) (LORENA et al., 2019), pode contribuir para enriquecer a análise dos dados. Na avaliação da qualidade do indicador composto IQCb se propõe a continuidade da pesquisa pela aplicação de testes de robustez (GRECO et al., 2018).

Finalmente, sugere-se o aprofundamento do estudo da relação entre Qualidade de Dados e Complexidade de Dados, incluindo novos indicadores para verificar o impacto nas relações descobertas neste trabalho. Além disso, se sugere pensar em outros trabalhos para validar a aplicação do modelo descoberto nesta pesquisa. O modelo resultante permite, entre outras coisas, saber a priori os tipos de problemas que um conjunto de dados apresenta e o provável desempenho do classificador caso nenhuma ação corretiva seja tomada. Nesse sentido, sugere-se pesquisar a utilização do modelo proposto para recomendar ações corretivas que reduzam o tempo de pré-processamento dos conjuntos de dados nas análises.

Referências

AGGARWAL, C. C. *Data mining: the textbook*. Springer International Publishing, 2015. Disponível em: <https://doi.org/10.1007/978-3-319-14142-8>. 47, 49, 60, 61

ANAGNOSTOPOULOS, I.; ZEADALLY, S.; EXPOSITO, E. Handling big data: research challenges and future directions. *The Journal of Supercomputing*, Springer, v. 72, n. 4, p. 1494–1516, 2016. Disponível em: <https://doi.org/10.1007/s11227-016-1677-z>. 18

AUER, F.; FELDERER, M. Addressing data quality problems with metamorphic data relations. In: *2019 IEEE/ACM 4th International Workshop on Metamorphic Testing (MET)*. IEEE, 2019. Disponível em: <https://doi.org/10.1109/met.2019.00019>. 23

AZEROUAL, O.; JHA, M. Without data quality, there is no data migration. *Big Data and Cognitive Computing*, v. 5, n. 2, 2021. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/5/2/24>. 22

BARELLA, V. H. et al. Data complexity measures for imbalanced classification tasks. In: IEEE. *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018. p. 1–8. Disponível em: <https://doi.org/10.1109/IJCNN.2018.8489661>. 19, 24, 36, 68

BERTI-EQUILLE, L. Data quality awareness: a case study for cost optimal association rule mining. *Knowledge and information Systems*, Springer, v. 11, n. 2, p. 191, 2007. Disponível em: <https://doi.org/10.1007/s10115-006-0006-x>. 18, 22, 46

BIALEK, W.; NEMENMAN, I.; TISHBY, N. Predictability, complexity, and learning. *Neural computation*, MIT Press, v. 13, n. 11, p. 2409–2463, 2001. Disponível em: <https://doi.org/10.1162/089976601753195969>. 35

BLAKE, R.; MANGIAMELI, P. The effects and interactions of data quality and problem complexity on classification. *Journal of Data and Information Quality (JDIQ)*, ACM New York, NY, USA, v. 2, n. 2, p. 1–28, 2011. 24, 59

BOSCHETTI, F. Mapping the complexity of ecological models. *ecological complexity*, Elsevier, v. 5, n. 1, p. 37–47, 2008. Disponível em: <https://doi.org/10.1016/j.ecocom-.2007.09.002>. 35, 63

BOSU, M. F.; MACDONELL, S. G. Experience: Quality benchmarking of datasets used in software effort estimation. *Journal of Data and Information Quality*, Association for Computing Machinery (ACM), v. 11, n. 4, p. 1–38, sep 2019. Disponível em: <https://doi.org/10.1145/3328746>. 22

BREIMAN, L.; CUTLER, A. *Random Forests*. [S.l.]: University of california, Berkeley, 2004. https://www.stat.berkeley.edu/~breiman/RandomForests/. Accessed February 09, 2021. 60

CASALICCHIO, G. et al. Openml: An r package to connect to the machine learning platform openml. *Computational Statistics*, Springer Nature, v. 32, n. 3, p. 1–15, 2017. Disponível em: <http://doi.acm.org/10.1007/s00180-017-0742-2>. 68, 97

CHEN, C. P.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, Elsevier, v. 275, p. 314–347, 2014. 17, 18

COUGO, P. *Modelagem conceitual e projeto de banco de dados*. [S.l.]: Elsevier Brasil, 2013. 50

DAVEY, A. et al. *Statistical power analysis with missing data: A structural equation modeling approach*. [S.l.]: Routledge, 2009. 52, 53

DIEBOLD, F. X. On the origin (s) and development of the term'big data'. PIER Working Paper, 2012. Disponível em: <http://dx.doi.org/10.2139/ssrn.2152421>. 17

FACELI, K. et al. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.]: Grupo Gen-LTC, 2000. 66

FANG, H. et al. A survey of big data research. *IEEE network*, NIH Public Access, v. 29, n. 5, p. 6, 2015. 17

FAYYAD, U. Data mining and knowledge discovery in databases: implications for scientific databases. In: IEEE. *Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No. 97TB100150)*. [S.l.], 1997. p. 2–11. 17

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Disponível em: <https://wvvw.aaai.org/ojs/index.php/aimagazine/article/download/1230/1131>. 17, 18, 19, 102

FEELDERS, A. Handling missing data in trees: Surrogate splits or statistical imputation? In: ŻYTKOW, J. M.; RAUCH, J. (Ed.). *Principles of Data Mining and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999. p. 329–334. ISBN 978-3-540-48247-5. 60

FERRARI, D. G.; SILVA, L. N. d. C. *Introdução a mineração de dados*. [S.l.]: Editora Saraiva, 2017. 18, 20, 47, 49, 52, 54, 59, 65

FOX, J. Getting started with the r commander: a basic-statistics graphical user interface to r. *J Stat Softw*, Citeseer, v. 14, n. 9, p. 1–42, 2005. 67, 70

GARCIA, L.; LORENA, A.; LEHMANN, J. Ecol: Complexity measures for classification problems. 2018. 19

GARCIA, L. et al. *ECoL: Complexity Measures for Supervised Problems*. [S.l.], 2020. R package version 0.4.0. Disponível em: <https://github.com/lpfgarcia/ECoL/>. 67, 68, 75, 79, 81, 82

GARCIA, L. P.; CARVALHO, A. C. de; LORENA, A. C. Effect of label noise in the complexity of classification problems. *Neurocomputing*, Elsevier, v. 160, p. 108–119, 2015. Disponível em: <https://doi.org/10.1016/j.neucom.2014.10.085>. 19, 23, 36, 68

GRECO, S. et al. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, Springer Science and Business Media LLC, v. 141, n. 1, p. 61–94, jan 2018. Disponível em: <https://doi.org/10.1007%2Fs11205-017-1832-9>. 24, 93, 102

HAIR, J. F. et al. *Multivariate Data Analysis*. [S.l.]: Pearson Education Limited, 2014. 10, 12, 22, 46, 50, 52, 54, 55, 56, 57, 58, 101

HAIR, J. F. et al. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 1. ed. [S.l.]: Sage publications, 2016. 12, 20, 25, 27, 31, 32, 33, 34, 35, 63, 65, 66, 84, 86, 88, 90, 91

HENSELER, J.; RINGLE, C.; SARSTEDT, M. Using partial least squares path modeling in advertising research: basic concepts and recent issues. In: _____. *Handbook of research on international advertising*. [S.l.]: Edward Elgar, 2012. p. 252–276. ISBN 9781848448582. 20, 29, 30, 33

HEY, A. J. et al. *The fourth paradigm: data-intensive scientific discovery*. [S.l.]: Microsoft research Redmond, WA, 2009. 17

HO, T. K.; BASU, M. Measuring the complexity of classification problems. In: IEEE. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*. [S.l.], 2000. v. 2, p. 43–47. 18, 19, 36, 68, 102

HO, T. K.; BASU, M. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, IEEE, n. 3, p. 289–300, 2002. Disponível em: <https://doi.org/10.1109/34.990132>. 19, 23, 35, 36, 68

HO, T. K.; BASU, M.; LAW, M. H. C. Measures of geometrical complexity in classification problems. In: *Data complexity in pattern recognition*. [S.l.]: Springer, 2006. p. 1–23. 23, 35

HäRDLE, W. K.; SIMAR, L. *Applied Multivariate Statistical Analysis*. 4. ed. Springer-Verlag Berlin Heidelberg, 2015. Disponível em: <https://doi.org/10.1007/978-3-662-45171-7>. 60

JAIN, A. *StatMeasures: Easy Data Manipulation, Data Quality and Statistical Checks*. [S.l.], 2015. R package version 1.0. Disponível em: <https://CRAN.R-project.org-/package=StatMeasures>. 67

JANUZAJ, E.; JANUZAJ, V. An application of data mining to identify data quality problems. In: *2009 Third International Conference on Advanced Engineering Computing and Applications in Sciences*. IEEE, 2009. Disponível em: <https://doi.org/10.1109-/advcomp.2009.11>. 22, 23

JAYAWARDENE, V.; SADIQ, S.; INDULSKA, M. An analysis of data quality dimensions. *ITEE Technical Report*, School of Information Technology and Electrical Engineering, The University of Queensland, v. 2015-02, p. 35–43, 2015. 18, 19, 20, 22, 46, 49, 63

KAMBATLA, K. et al. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, Elsevier, v. 74, n. 7, p. 2561–2573, 2014. 17

KARKOUCH, A. et al. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, Elsevier BV, v. 73, p. 57–81, sep 2016. Disponível em: <https://doi.org/10.1016/j.jnca.2016.08.002>. 22

LARANJEIRO, N.; SOYDEMIR, S. N.; BERNARDINO, J. A survey on data quality: Classifying poor data. In: *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2015. Disponível em: <https://doi.org/10.1109-/prdc.2015.41>. 22

LAROSE, D. T. *Data mining and predictive analytics*. 2. ed. [S.l.]: John Wiley & Sons, 2015. 18

LATAN, H.; NOONAN, R. *Partial least squares path modeling: Basic concepts, methodological issues and applications*. Springer, 2017. Disponível em: <https://doi.org-/10.1007/978-3-319-64069-3>. 29

LIBóRIO, M. P. et al. Measuring intra-urban inequality with structural equation modeling: A theory-grounded indicator. *Sustainability*, v. 12, n. 20, 2020. ISSN 2071-1050. Disponível em: <https://www.mdpi.com/2071-1050/12/20/8610>. 24, 94, 95

LIEBENAU, J.; BACKHOUSE, J. *Understanding Information*. Macmillan Education UK, 1990. Disponível em: <https://doi.org/10.1007/978-1-349-11948-6>. 22

LORENA, A. C.; CARVALHO, A. C. de. Evaluation of noise reduction techniques in the splice junction recognition problem. *Genetics and Molecular Biology*, SciELO Brasil, v. 27, n. 4, p. 665–672, 2004. Disponível em: <https://doi.org/10.1590/S1415-47572004000400031>. 68

LORENA, A. C. et al. Analysis of complexity indices for classification problems: Cancer gene expression data. *Neurocomputing*, Elsevier, v. 75, n. 1, p. 33–42, 2012. Disponível em: <https://doi.org/10.1016/j.neucom.2011.03.054>. 45

LORENA, A. C. et al. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 5, p. 1–34, 2019. Disponível em: <https://doi.org/10.1145/3347711>. 12, 36, 37, 38, 39, 40, 42, 43, 45, 65, 67, 68, 69, 75, 77, 79, 80, 82, 88, 102

LORENA, A. C.; SOUTO, M. C. de. On measuring the complexity of classification problems. In: SPRINGER. *International Conference on Neural Information Processing*. 2015. p. 158–167. Disponível em: <https://doi.org/10.1007/978-3-319-26532-2_18>. 36

MCKNIGHT, P. E. et al. *Missing data: A gentle introduction*. [S.l.]: Guilford Press, 2007. 10, 12, 50, 51, 52, 53, 54, 55, 66

NARDO, M. et al. *Handbook on constructing composite indicators: methodology and user guide*. [S.l.]: OECD publishing, 2008. 24, 93, 94, 95, 96

NISBET, R.; ELDER, J.; MINER, G. *Handbook of statistical analysis and data mining applications*. [S.l.]: Academic Press, 2009. 60

PROJECT, R. *Using imputation for missing values*. [S.l.]: National Centre for Research Methods, Economic and Social Research Council, 2009. `https://www.restore.ac.uk/PEAS/imputation.php/`. Accessed January 11, 2021. 54

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014. 60

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. 67

RINGLE, C. M.; WENDE, S.; BECKER, J.-M. Smartpls 3. *Boenningstedt: SmartPLS GmbH, http://www. smartpls. com*, 2015. 84

ROSLI, M. M.; TEMPERO, E.; LUXTON-REILLY, A. Can we trust our results? a mapping study on data quality. In: *2013 20th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2013. Disponível em: <https://doi.org/10.1109/apsec.2013-.26>. 22

ROUSSEEUW, P. J.; ZOMEREN, B. C. V. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, Taylor & Francis, v. 85, n. 411, p. 633–639, 1990. 49

RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976. 10, 52, 53, 54, 101

SÁNCHEZ, J. S.; MOLLINEDA, R. A.; SOTOCA, J. M. An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis and Applications*, Springer, v. 10, n. 3, p. 189–201, 2007. Disponível em: <https://doi.org/10.1007/s10044-007-0061-2>. 19, 23, 36, 68

SARSTEDT, M.; RINGLE, C. M.; HAIR, J. F. Partial least squares structural equation modeling. *Handbook of market research*, Springer Heidelberg, v. 26, p. 1–40, 2017. 10, 25, 26, 27, 28, 31, 34

SILVA, L. A. d.; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. 1. ed. [S.l.]: Elsevier, 2016. 18, 20, 47, 54, 59, 65

SING, T. et al. Rocr: visualizing classifier performance in r. *Bioinformatics*, Oxford University Press, v. 21, n. 20, p. 3940–3941, 2005. 68

SOPER, D. S. *A-priori sample size calculator for structural equation models [Software]*. 2017. https://www.danielsoper.com/statcalc/calculator.aspx?id=1. Accessed February 17, 2021. 66

STATISTICS, U. D. o. L. Bureau of L. *New CPI estimator expected to lower inflation rate by 0.2 percent*. 1999. https://www.bls.gov/opub/ted/1999/Mar/wk4/art03.htm. 96

STREUKENS, S.; LEROI-WERELDS, S. Bootstrapping and pls-sem: A step-by-step guide to get more out of your bootstrap results. *European Management Journal*, Elsevier, v. 34, n. 6, p. 618–632, 2016. Disponível em: <https://doi.org/10.1016/j.emj.2016.06-.003>. 33

TALEB, I. et al. Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data*, Springer Science and Business Media LLC, v. 8, n. 1, may 2021. Disponível em: <https://doi.org/10.1186/s40537-021-00468-0>. 22, 23

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. 2. ed. [S.l.]: Pearson Education India, 2018. 18, 47, 59, 60, 61

TENENHAUS, M. et al. Pls path modeling. *Computational statistics & data analysis*, Elsevier, v. 48, n. 1, p. 159–205, 2005. 20, 25

TENG, D. et al. Vdqam: A toolkit for database quality evaluation based on visual morphology. In: IEEE. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. [S.l.], 2012. p. 245–246. 20

TIWARI, V.; KASHIKAR, A. *OutlierDetection: Outlier Detection*. [S.l.], 2019. R package version 0.1.1. Disponível em: <https://CRAN.R-project.org-/package=OutlierDetection>. 67

TOMASELLI, V.; FORDELLONE, M.; VICHI, M. Building well-being composite indicator for micro-territorial areas through PLS-SEM and k-means approach. *Social Indicators Research*, Springer Science and Business Media LLC, v. 153, n. 2, p. 407–429, aug 2020. Disponível em: <https://doi.org/10.1007%2Fs11205-020-02454-0>. 24, 94

VALVERDE, M. C. et al. Applying a data quality model to experiments in software engineering. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2014. p. 168–177. Disponível em: <https://doi.org/10.1007/F978-3-319-12256-4_18>. 22

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, Informa UK Limited, v. 12, n. 4, p. 5–33, mar 1996. Disponível em: <https://doi.org/10.1080/07421222.1996.11518099>. 22

WOOK, M. et al. Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling. *Journal of Big Data*, v. 8, 2021. Disponível em: <https://doi.org/10.1186/s40537-021-00439-5>. 22

YE, N.; CHEN, Q. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, Wiley Online Library, v. 17, n. 2, p. 105–112, 2001. 49

ZUBEK, J.; PLEWCZYNSKI, D. M. Complexity curve: a graphical measure of data complexity and classifier performance. *PeerJ Computer Science*, PeerJ Inc., v. 2, p. e76, 2016. Disponível em: <https://doi.org/10.7717/peerj-cs.76>. 19, 24, 36, 68

ZWICKER, R.; SOUZA, C. A. d.; BIDO, D. d. S. Uma revisão do modelo do grau de informatização de empresas: novas propostas de estimação e modelagem usando pls (partial least aquares). *Anais. ENANPAD - Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração*, 2008. 20, 32