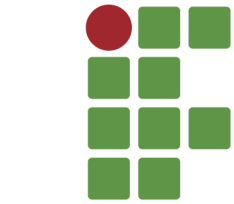


INSTITUTO FEDERAL
Brasília

Instituto Federal de Brasília
Bacharelado em Ciência da Computação
Campus Taguatinga
http://computacaoifb.net

ANÁLISE DA QUALIDADE DOS DADOS DO SISTEMA NACIONAL DE INFORMAÇÕES DA EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA DO INSTITUTO FEDERAL DE BRASÍLIA

Por

NÍNIVE HELEN HORÁCIO DA SILVA

Trabalho de Graduação

BRASÍLIA/2025

Ficha de identificação da obra elaborada pelo bibliotecário
Marcelo José Rodrigues da Conceição (CRB1-2323)

Silva, Nínive Helen Horácio da

S586a Análise da qualidade dos dados do Sistema Nacional de Informações da Educação Profissional e Tecnológica do Instituto Federal de Brasília / Nínive Helen Horácio da Silva. Brasília-DF, 2025.

62 f. : il.

Trabalho de Conclusão de Curso de Bacharelado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia de Brasília, Campus Taguatinga, 2025.

Orientador: Prof. Dr. Fábio Henrique Monteiro Oliveira.

Inclui referências.

1. Estruturas de dados (Computação). 2. Processamento eletrônico de dados. 3. Análise de dados. 4. Big data. 5. Linguagem de programação (Computadores). I. Título. II. Oliveira, Fábio Henrique Monteiro. III. Instituto Federal de Educação, Ciência e Tecnologia de Brasília.

CDU 004.6

Trabalho de Graduação apresentado por **Nínive Helen Horácio da Silva** ao programa de Graduação em Ciência da Computação do Instituto Federal de Brasília, sob o título **Análise da qualidade dos dados do Sistema Nacional de Informações da Educação Profissional e Tecnológica do Instituto Federal de Brasília**, orientada pelo **Prof. Dr. Fábio Henrique Monteiro Oliveira** e aprovada pela banca examinadora formada pelos avaliadores:

Prof. Me. João Victor de Araújo Oliveira
Computação/IFB

Patrícia Rodrigues Amorim
Técnica em Assuntos Educacionais
Coordenadora-Geral de Planejamento /IFB

BRASÍLIA
2025

Resumo

O Conjunto de dados maior e mais complexo (Big Data) está presente em diversos setores públicos e privados, onde grandes volumes de dados são armazenados e usados para apoiar a tomada de decisões. Manter a qualidade e a integridade desses dados é um desafio, pois eles precisam seguir regras de negócio específicas para garantir que as informações sejam confiáveis. Segundo o grupo *Total Data Quality Management* do MIT, liderado pelo professor Richard Y. Wang, a qualidade dos dados é definida como "adequação para o uso", considerando dimensões como exatidão, completude, integridade, unicidade, consistência, entre outras. No Brasil, o Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec) coleta e armazena dados sobre cursos técnicos e tecnológicos. Este trabalho teve como objetivo validar os dados de matrículas do Sistec referentes ao Instituto Federal de Brasília (IFB), verificando se estão em conformidade com as regras de negócio e analisando a qualidade dos dados por meio de métricas específicas. Para isso, foram desenvolvidos scripts em Python que aplicaram as validações e mensuraram as dimensões da qualidade dos dados. Os resultados mostraram que, embora algumas dimensões apresentem boa qualidade, a consistência dos dados precisa ser melhorada, pois concentra a maioria das inconsistências. Após a análise, são sugeridas correções para reduzir as irregularidades nos dados, o que pode aumentar a qualidade das informações extraídas e evitar perdas causadas pela má qualidade.

Palavras-chave: Big Data, Governança de Dados, Qualidade de Dados, Análise de Qualidade de Dados, Sistec, IFB

Abstract

Larger and more complex datasets (Big Data) are present across various public and private sectors, where vast volumes of data are stored and used to support decision-making. Ensuring the quality and integrity of this data is a challenge, as it must follow specific business rules to ensure the reliability of the information. According to the Total Data Quality Management group at the Massachusetts Institute of Technology (MIT), led by Professor Richard Y. Wang, data quality is defined as "fitness for use," considering dimensions such as accuracy, completeness, integrity, uniqueness, consistency, among others. In Brazil, the Sistema Nacional de Informações da Educação Profissional e Tecnológica (National System of Information on Professional and Technological Education – Sistec) collects and stores data on technical and technological courses. This study aimed to validate enrollment data from Sistec related to the Instituto Federal de Brasília (Federal Institute of Brasília – IFB), verifying their compliance with business rules and analyzing data quality through specific metrics. To achieve this, Python scripts were developed to apply validations and measure data quality dimensions. The results showed that, although some dimensions demonstrated good quality, data consistency needs improvement, as it concentrates most of the inconsistencies. Based on the analysis, corrections are suggested to reduce data irregularities, which may improve the quality of extracted information and help prevent losses caused by poor data quality.

Keywords: Big Data, Data Governance, Data Quality, Data Quality Analysis, Sistec, IFB

Agradecimentos

Primeiramente, agradeço a Deus e à fé que me manteve firme durante todos esses anos de faculdade. Também sou grata à minha família e amigos, pois o apoio deles foi fundamental para minha trajetória. Agradeço ao IFB pela excelente educação pública e gratuita, assim como a todos os professores que dedicam seu tempo ao nosso aprendizado, e agradeço ao meu orientador por ter me guiado na realização deste trabalho de conclusão de curso. Estendo minha gratidão a todos os servidores que trabalham no IFB, contribuindo para o bom funcionamento da instituição. Agradeço ainda pela oportunidade que tive de participar do projeto de pesquisa Termo de Execução Descentralizada (TED) 10718/2021, cujo tema foi "A definição de métodos e procedimentos para aplicação de higienização". Um dos meus objetivos ao estudar em uma instituição federal era participar de projetos de pesquisa, pois acredito que a pesquisa agrega valor tanto ao estudante quanto à sociedade. Também agradeço a todos os locais onde estagiei e trabalhei durante a graduação, que fizeram parte importante do meu aprendizado.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

Sumário

1
Introdução

Big Data lida com dados em grande escala e de diversas fontes. ZAKIR; SEYMOUR; BERG (2015) descrevem o *Big Data* como dados derivados de várias fontes, que envolvem diversos paradigmas estruturados e não estruturados. Esses dados podem ser extraídos e contêm informações importantes, que podem ser utilizadas para uma maior percepção nas tomadas de decisões, utilizando métodos e tecnologias de apoio; uma delas é a Ciência de Dados (CD).

A CD utiliza bases de *Big Data* para extrair e analisar informações e é um campo emergente que se concentra em extrair informações relevantes de conjuntos de dados complexos. Esse método possibilita o estudo dos dados, aplicando uma ciência para extrair informações que serão aplicadas para melhorias em vários setores, tanto públicos quanto privados (DANIEL, 2018).

Com a complexidade de manipular grandes bases de dados e extrair informações, tornou-se necessário o aprimoramento de regras para a administração desses dados, que vão além do controle de permissões para acesso à informação ou à extração e manipulação dos dados. Os dados possuem uma variedade de elementos, os quais necessitam de métodos para promover a qualidade dos dados e para o gerenciamento na extração de informações para a tomada de decisões.

Sendo assim, surgiram as áreas de Governança de Dados (GD) e Qualidade de Dados (QD). A GD é um conjunto de processos, padrões e estruturas que garantem uma gestão adequada dos dados em organizações. Já a QD é uma forma de avaliar a qualidade dos dados já armazenados, garantindo sua integridade e aplicando melhorias na qualidade dos dados.

O setor público é um dos que estão investindo na área de governança de dados e na utilização de informações para a tomada de decisões. Com o avanço da tecnologia, surgiu o Governo Digital, que modernizou a administração do Estado brasileiro. Esse programa utiliza dados disponíveis para otimizar e transformar os serviços públicos (Tribunal de Contas da União, 2024). O governo está aumentando a utilização de dados para obter informações a partir de várias bases de dados diferentes, a fim de aplicar medidas públicas de maneira inteligente. Este projeto consiste em instituir e disponibilizar, por meio de compartilhamento, um conjunto de dados a ser utilizado pelos órgãos como referência para a execução de sua gestão, formulação de políticas ou para a oferta de serviços públicos (Governo Digital, 2024).

Uma das implementações sobre a coleta de dados no setor público do Brasil foi a criação do Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec). Esse sistema tem como finalidade servir como um mecanismo de registro e divulgação dos dados da Educação Profissional e Tecnológica (EPT) e como instrumento de validação e expedição de diplomas de cursos de educação profissional técnica de nível médio. O Sistec foi instituído e implantado pelo Ministério da Educação (MEC) em 2009, com a intenção de fornecer ao governo dados que permitissem a elaboração de indicadores educacionais confiáveis, desenvolvendo e fortalecendo a educação profissional e tecnológica nacional (LIMA MACHADO, 2019). O Sistec também auxilia o Programa Nacional de Acesso ao Ensino Técnico e Emprego (Pronatec), que tem a finalidade de ampliar a oferta de cursos de EPT por meio de programas, projetos e

ações de assistência técnica e financeira (Ministério da Educação, 2024a).

1.1 Problema

Ao lidar com dados, um dos desafios é a qualidade das informações armazenadas, para extrair dados e realizar análises que possibilitem a tomada de decisões. Segundo TALEB et al. (2016), inicialmente os dados são incompletos e podem conter muitas inconsistências, como dados errados, faltantes ou incompletos.

Essas irregularidades podem ser causadas por vários fatores, principalmente ao lidar com grandes volumes de dados, como os de *Big Data*. Em ambientes de *Big Data*, os dados são o elemento mais importante por serem utilizados para o processamento e análise; porém, sem dados adequados para uso, essas fases não terão sucesso. Dados faltantes ou de baixa qualidade podem comprometer a eficácia das análises e da tomada de decisões, impossibilitando a obtenção de resultados precisos e confiáveis.

Apenas possuir um grande volume de dados não é suficiente; é necessário garantir a qualidade desses dados, pois etapas de governança de dados às vezes são negligenciadas, fazendo com que haja um grande volume de dados com baixa qualidade. Para garantir a qualidade dos dados, é necessário utilizar métodos de governança de dados, com regras bem definidas para a coleta e armazenamento. Essas regras estabelecem padrões para o formato, valores e estruturas dos dados, assegurando que os dados sejam consistentes e estejam no formato correto desde a coleta até o armazenamento.

Quando os dados não seguem as regras de negócio, podem surgir inconsistências que comprometem a qualidade das informações. Para lidar com essas inconsistências, é essencial aplicar metodologias de qualidade dos dados para validar se eles estão conforme as regras de negócio e se são adequados para uso. Essas regras devem ser aplicadas não apenas para garantir que os dados estejam corretos, mas também para avaliar sua qualidade.

1.2 Proposta

Como mencionado, possuir um grande volume de dados não é suficiente; é essencial que os dados tenham qualidade para que seja possível extrair informações confiáveis e cumprir seus objetivos. Dado que a qualidade dos dados é fundamental, este trabalho propõe analisar a qualidade dos dados do Sistec de registros de matrículas Instituto Federal de Brasília (IFB) disponíveis no portal de dados abertos do IFB. A análise inclui a validação desses dados com base nas regras de negócio previamente definidas pelo projeto de pesquisa mencionado na seção justificativa e, após a validação, a identificação de inconsistências.

Propõe-se realizar uma análise detalhada das irregularidades, investigando cada caso de forma qualitativa e quantitativa para avaliar seu impacto nas informações. Essa investigação será feita individualmente, aplicando regras de negócio específicas para identificar quais dados não estão em conformidade e registrando as inconsistências, caso existam. Por exemplo, se uma regra de negócio define que um campo de data deve ser preenchido em um formato específico, registros que não seguirem esse padrão terão suas inconsistências relatadas.

Dessa forma, será possível mensurar o grau de perda de informação causado por essas irregularidades, avaliando como elas podem comprometer a utilidade dos dados e a extração de informações relevantes. Ao identificar os registros afetados, serão aplicadas metodologias de qualidade de dados para determinar quais dimensões foram comprometidas. Algumas dimensões a serem analisadas são:

- Exatidão: refere-se ao grau em que os dados estão próximos do valor verdadeiro ou correto, conhecido como referência. Quanto maior a exatidão, mais os dados refletem fielmente a realidade ou o valor esperado.
- Completude: verifica se todos os dados necessários estão presentes.
- Consistência: refere-se às regras semânticas definidas para os dados, conforme a teoria relacional. Isso implica que os dados devem seguir regras de restrição predefinidas durante o cadastro da informação.

As dimensões mencionadas acima são algumas das que devem ser analisadas. Após identificar e mensurar a qualidade dos dados, serão propostas sugestões de melhorias para aumentar a qualidade dos dados e reduzir a ocorrência de futuras inconsistências. Essas sugestões têm como objetivo aprimorar a qualidade das informações coletadas e prevenir a repetição de irregularidades nos dados. As recomendações serão elaboradas com base nas dimensões afetadas, garantindo que as ações corretivas sejam direcionadas à melhoria da qualidade dos dados.

1.3 Justificativa

Alguns estudos realizados pela equipe do projeto de pesquisa, em parceria com a Secretaria de Educação Profissional e Tecnológica Sistec e autorizados pelo Termo de Execução Descentralizada (TED), têm como uma de suas finalidades a execução de programas, projetos e atividades de interesse recíproco (Conselho Nacional de Desenvolvimento Científico e Tecnológico, 2024). O projeto foi realizado por uma equipe do IFB, com base no TED 10718/2021, e tem o tema: "A definição de métodos e procedimentos para aplicação de higienização"(A higienização de dados é o processo de identificar, corrigir ou remover dados incorretos) (SIMEC, 2024).

Durante o projeto citado, foi utilizada a base de dados do Sistec, que contém registros da educação profissional e tecnológica do Brasil, para criar validações nos registros e relatar as inconsistências encontradas. Além disso, foi desenvolvido um painel para visualização das informações (OLIVEIRA et al., 2024). Nos estudos realizados pela equipe, foram conduzidas etapas de documentação das regras de negócio e criação de funções para validação dos dados, aplicando as regras de negócio. Após as validações, foi realizada uma análise quantitativa da quantidade de inconsistências identificadas nos registros.

Inspirado no projeto mencionado, este trabalho de conclusão de curso visa analisar as inconsistências dos dados de forma qualitativa e quantitativa, com foco na avaliação da qualidade dos dados e na proposição de sugestões para minimizar o surgimento de novas inconsistências. Para realizar essa análise, serão aplicadas metodologias de qualidade de dados, conforme detalhado na seção de Proposta

Neste projeto, os dados utilizados são provenientes do Sistec do IFB, estão disponíveis no portal de dados abertos do IFB. O foco principal é analisar e validar os dados do Sistec que contém registros do IFB, realizando uma análise qualitativa para propor sugestões de melhorias na qualidade dos dados.

1.4 Objetivos

1.4.1 Objetivo Geral

O objetivo deste trabalho de conclusão de curso é realizar uma análise quantitativa e qualitativa da qualidade dos dados do Sistec provenientes do IFB, identificando inconsistências e

propondo estratégias de melhoria, com o propósito de aumentar a confiabilidade das informações.

1.4.2 Objetivos Específicos

- Analisar a qualidade dos dados do Sistec do IFB, utilizando as funções desenvolvidas pela equipe do projeto no âmbito do TED.
- Assegurar que os registros do sistema estejam corretos, validando-os conforme as regras de negócio estabelecidas.
- Identificar e registrar todas as inconsistências encontradas nos dados, para compreender onde ocorrem irregularidades.
- Categorizar as inconsistências de acordo com as dimensões de qualidade impactadas, para orientar as correções necessárias.
- Avaliar o impacto dessas inconsistências aplicando métricas específicas, entendendo como elas afetam a qualidade dos dados.
- Sugerir melhorias nos processos de coleta e gestão dos dados, com base nos resultados da análise, para aumentar a confiabilidade e a integridade das informações.

2
Revisão da Literatura

A revisão da literatura é uma parte importante do processo de investigação. ela envolve localizar, analisar, sintetizar e interpretar pesquisas prévias (como revistas científicas, livros, atas de congressos, resumos, etc.) (BENTO, 2012). Nessa etapa, são selecionados materiais semelhantes em algum ponto ao trabalho a ser desenvolvido, permitindo uma análise crítica das pesquisas existentes. Essa fase contribui para identificar práticas, metodologias e teorias aplicáveis, fornecendo embasamento para o desenvolvimento do trabalho.

Para este trabalho de conclusão de curso, foi utilizado como orientação para a revisão da literatura o *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)*, que oferece um guia para realizar revisões e meta-análises de trabalhos científicos. Esse método fornece estratégias para a seleção e avaliação dos materiais, garantindo que a revisão seja abrangente e bem fundamentada.

2.1 Método PRISMA

O método *PRISMA* orienta as etapas da revisão da literatura, facilitando a busca por trabalhos de pesquisa relacionados ao tema em questão. Um dos métodos mais comuns para formular questões de investigação dentro dessa abordagem é o modelo População, Intervenção, Comparação, *Outcomes* (Desfecho) (PICO). População: qual população está sendo considerada. Intervenção: qual intervenção está sendo analisada. Comparação: com o que a intervenção será comparada. *Outcomes* (Desfecho): qual desfecho se planeja avaliar. (DONATO; DONATO, 2019). O quadro 2.1 apresenta uma descrição detalhada de cada elemento do acrônimo PICO.

Quadro 2.1: Questões de investigação usando o método PICO

Acrônimo	Definição
P	Dados sobre informações de pessoas/instituições.
I	Investigar como os dados inconsistentes podem afetar as informações.
C	Comparar o impacto das inconsistências nos dados sobre as informações.
O	Artigos que contêm análise de dados, análise quantitativa e qualitativa, investigando inconsistências nos dados e mostrando resultados sobre o impacto das irregularidades nas informações.

Fonte: Elaborado pela autora

A seleção dos artigos foi realizada nas bases de dados da Coordenação de Aperfeiço-

amento de Pessoal de Nível Superior (CAPES) e do Google Acadêmico, baseado no método *PRISMA* para a seleção dos artigos. Na base da CAPES, a pesquisa foi feita na seção de acervo, utilizando a opção de busca por assunto. Já no Google Acadêmico, a pesquisa foi realizada diretamente na barra de pesquisa.

Após definir os métodos de investigação, as palavras-chave foram selecionadas por assunto e utilizadas na busca dos artigos para a revisão sistemática. Para a pesquisa na CAPES, as palavras-chave foram: *Data Quality Assessment AND Data Quality Assessment Dimensions AND Big Data AND Data Quality AND Big Data Quality.*. Já para a pesquisa no Google Acadêmico, o termo utilizado foi: *Big Data Quality*.

Para a revisão da literatura, foram selecionados 4 artigos científicos publicados entre 2016 e 2024, seguindo os critérios de inclusão e exclusão, os quais estão detalhados no quadro 2.2.

Quadro 2.2: Critérios de inclusão e exclusão dos artigos para a revisão da literatura

Critérios de inclusão	**Critérios de exclusão**
Artigos que apresentam validação de dados e avaliam a qualidade dos dados através das inconsistências relatadas durante a validação, contendo etapas de processamento de dados, validação e analise da qualidade.	Artigos que não citam a qualidade dos dados e sua importância para a qualidade das informações e que não apresentam validação de dados utilizando as regras de negócio dos dados.
Artigos que apresentam métodos para analisar a qualidade dos dados.	Artigos que não apresentam métodos de qualidade dos dados e não citam inconsistências de dados.
Artigos que apresentam métodos de qualidade dos dados e métricas, fórmulas para medir a qualidade dos dados.	Artigos que não apresentam as dimensões e métricas para medir a qualidade dos dados.
Artigos que apresentam análises de qualidade dos dados e dimensões sobre como medir a qualidade dos dados, além de apresentar resultados sobre o impacto da má qualidade dos dados e sua importância.	Artigos que não apresentam nenhuma análise sobre a qualidade dos dados e resultados da análise ou importância de uma análise de qualidade nos dados

Fonte: Elaborado pela autora

Após definir os critérios de pesquisa, foram estabelecidos filtros, incluindo a seleção de artigos de acesso aberto e publicados entre os anos de 2016 e 2024, a escolha do intervalo foi escolhida de maneira que tivesse o resultado de grande quantidade de artigos. Na pesquisa realizada na plataforma da CAPES, foram encontrados 339 artigos, desses 339 artigos encontrados, foi feito uma procurar rápida se tinha termos como *data quality*, *metrics*, *data quality metrics* que se referem a qualidade dos dados e métricas de qualidade de dados, pesquisando dessa forma foi lido o resumo de 20 artigos, Após a leitura dos resumos de 20 artigos, 8 foram selecionados para uma leitura mais detalhada, seguindo os critérios de inclusão e exclusão. Com a leitura detalhada desses 8 artigos, 3 foram incluídos, enquanto 5 foram excluídos. Os 3 artigos incluídos foram selecionados para a revisão da literatura.

Na pesquisa realizada no Google Acadêmico, após a leitura dos resumos dos quatro primeiros artigos encontrados, esses foram selecionados para uma análise mais detalhada, conforme os critérios de inclusão e exclusão estabelecidos. Após a leitura completa, um artigo foi incluído e três foram excluídos. Foram selecionados artigos nas plataformas CAPES e Google Acadêmico, e elaborados resumos de quatro artigos para a revisão da literatura.

2.2 Trabalhos Analisados

Esta seção apresenta a extração e síntese dos dados dos artigos selecionados, baseados no método o método *PRISMA*.

2.2.1 Determinants of Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework

Este trabalho realiza uma análise sobre a qualidade dos dados relacionados às rodovias. Conforme discutido por KRISHNA; RUIKAR; JHA (2023), as agências rodoviárias coletam uma grande quantidade de dados, que variam desde informações preliminares de pesquisa até dados detalhados sobre as condições do pavimento. Os autores também citam o relatório do Fundo Monetário Internacional (FMI) de 2019. A FMI visa garantir a estabilidade econômica e financeira global, promovendo o crescimento econômico dos países-membros. O relatório citado, intitulado *Big Data Equals Big Questions for the Engineering and Construction Industry*, diz que alguns dos projetos de infraestrutura mais significativos demandam, em média, 130 milhões de e-mails, 55 milhões de documentos e 12 milhões de fluxos de trabalho. No entanto, 95,5% de todos os dados coletados na indústria de engenharia e construção não são utilizados devido à dificuldade que muitas empresas enfrentam para gerenciar e processar esse grande volume de informações.

Os autores também mencionam um relatório da indústria de 2018, intitulado *Construction Disconnected*, da FMI. Nesse relatório, observa-se que 48% de todos os retrabalhos em projetos de infraestrutura nos Estados Unidos são causados por dados inadequados e comunicação incorreta. Globalmente, uma média de 52% dos retrabalhos é atribuída a problemas de dados e comunicação, sendo que 34,4% dos retrabalhos são causados por dados incorretos, desatualizados ou falhos, enquanto 28,8% resultam da dificuldade de acessar os dados necessários.

Os autores, em seu trabalho, investigaram a qualidade dos dados de infraestrutura rodoviária para avaliar a exatidão das informações, destacando que a má qualidade dos dados afeta negativamente a tomada de decisões em projetos de infraestrutura rodoviária. De acordo com KRISHNA; RUIKAR; JHA (2023, p. 3), "Avaliar a qualidade dos dados é fundamental para as organizações e destaca a importância de identificar as dimensões que definem a qualidade dos dados". Para analisar a qualidade dos dados das rodovias, os autores dividiram o estudo em três objetivos principais:

- Estabelecer as dimensões de qualidade dos dados necessárias para avaliar a qualidade das informações sobre as estruturas rodoviárias, facilitando a tomada de decisões.
- Determinar a importância de cada dimensão da qualidade dos dados em diferentes níveis de tomada de decisão.
- Definir a prioridade das dimensões nas categorias da estrutura semiótica.

Para atingir os objetivos propostos, os autores do artigo *Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework* aplicaram duas metodologias

de avaliação da qualidade dos dados: a estrutura de avaliação da qualidade dos dados e o *framework* semiótico, *framework* Estrutura ou conjunto de ferramentas que oferecem suporte para a construção de sistemas ou processos .

A estrutura de avaliação da qualidade dos dados é utilizada para identificar as dimensões da qualidade, sendo que cada dimensão possui uma categoria associada, o que permite diferenciar e mensurar a qualidade de cada uma. As principais dimensões analisadas incluem exatidão, completude, consistência e relevância, entre outras, e cada uma está associada a atributos mensuráveis. Essas categorias foram definidas com base nas necessidades específicas de agências rodoviárias e de engenharia.

O *framework* semiótico é aplicado para estruturar e avaliar a qualidade dos dados sobre infraestruturas rodoviárias em diferentes níveis de decisão, garantindo que os dados estejam em um formato que permita decisões confiáveis em cada etapa do projeto de rodovias. Esse *framework* segue etapas citadas pelos autores: nível empírico, nível sintático, nível semântico e nível pragmático.

- Nível empírico: analisa a confiabilidade dos dados.
- Nível sintático: examina a estrutura dos dados para verificar se estão organizados adequadamente para análise.
- Nível semântico: foca na interpretação dos dados, garantindo que representem com exatidão as condições das rodovias.
- Nível pragmático: avalia a aplicação prática dos dados nas decisões de infraestrutura, como a escolha de tratamentos para pavimentos danificados.

Após definir estratégias para analisar a qualidade dos dados de rodovias, os autores aplicaram três etapas. Primeiro, identificaram as dimensões de qualidade dos dados da infraestrutura rodoviária usando o *framework* semiótico, aplicando as dimensões apropriadas ao projeto. Na segunda etapa, foi realizado um questionário para avaliar as dimensões da qualidade dos dados selecionadas na primeira etapa. Esse questionário foi criado com base nas dimensões escolhidas e teve como objetivo coletar informações das partes interessadas na infraestrutura rodoviária. A estrutura semiótica consiste em 43 dimensões de qualidade dos dados; para a análise da qualidade dos dados de projetos de infraestrutura rodoviária, foram selecionadas 20 dimensões. O questionário, focado nessas 20 dimensões de qualidade dos dados, foi validado por três especialistas em projetos rodoviários e passou por um estudo piloto para ajustar sua clareza. Ele foi distribuído a 220 partes interessadas da *National Highway Authority of India (NHAI)*, resultando em 105 especialistas que participaram da pesquisa, representando uma taxa de resposta de 48%.

O estudo conclui que as cinco principais dimensões para garantir a integridade dos dados são: exatidão, acessibilidade, integridade, consistência e pontualidade. Além dessas, dimensões como relevância, interpretabilidade e credibilidade também são importantes. A análise das dimensões revela que, para uma tomada de decisão eficaz, é essencial abordar cada uma delas de forma sistemática. Essa abordagem contribui para a qualidade geral dos dados em projetos de infraestrutura rodoviária, melhorando a confiabilidade das decisões e promovendo um uso mais eficiente dos recursos, resultando na realização mais eficaz dos objetivos do projeto.

Os autores apresentam as categorias das dimensões selecionadas e os resultados das métricas utilizadas para calcular a qualidade das informações em forma de tabela. A tabela mostra a importância das dimensões de qualidade dos dados em cada etapa da tomada de decisão, como nos níveis de decisão estratégica, de rede, de programa, de seleção de projeto e de execução

de projetos de infraestrutura rodoviária, com as principais dimensões de qualidade dos dados e os resultados das métricas das categorias das dimensões.

Os resultados da avaliação da qualidade dos dados sobre a infraestrutura rodoviária destacam a importância de um sistema estruturado de avaliação da qualidade dos dados. Esse sistema não apenas aumenta a eficácia das decisões relacionadas a projetos de infraestrutura rodoviária, mas também é essencial para garantir a sustentabilidade a longo prazo. Dimensões como exatidão, consistência e integridade dos dados desempenham um papel fundamental nas tomadas de decisão baseadas em dados, reduzindo erros e resultando em dados mais eficazes. Portanto, é necessário que as organizações implementem práticas de qualidade dos dados.

2.2.2 An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis

A capacidade de coletar e processar grandes volumes de dados possibilitou que as organizações extraíssem informações para aprimorar a tomada de decisões. No artigo *An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis*, os autores afirmam que "a análise de *Big Data* se tornou um componente fundamental da tomada de decisões, permitindo que as organizações revelem padrões, tendências e correlações ocultas que antes eram inacessíveis"(ELOUATAOUI; EL MENDILI; GAHI, 2023, p. 1). No entanto, é essencial que os dados de *Big Data* apresentem alta qualidade, pois isso desempenha um papel crucial na exatidão e confiabilidade das análises extraídas. De acordo com uma pesquisa da Gartner, a baixa qualidade dos dados gera um gasto médio de US$ 12,9 milhões anuais, além de impactar negativamente a receita a longo prazo. Dados de baixa qualidade aumentam a complexidade e contribuem para decisões equivocadas.

A proposta dos autores é desenvolver um *framework* automatizado que integra várias etapas para detectar e corrigir anomalias nos dados. O modelo é baseado no *Extreme Gradient Boosting (XGBoost)*, um algoritmo de *Aprendizado de maquina (Machine Learning)* especializado em problemas de classificação e regressão. O *framework* foca em uma estrutura mais abrangente, oferecendo uma solução para todas as dimensões afetadas, e não apenas para contextos específicos. Os autores destacam três principais contribuições para a correção dessas irregularidades:

- Contribuição 1: uma estrutura sofisticada para a correção de inconsistências e anomalias, baseada em um modelo preditivo, capaz de detectar e corrigir inconsistências difíceis.
- Contribuição 2: abordagem para lidar com as seis dimensões críticas que afetam a qualidade dos dados: exatidão, completude, conformidade, unicidade, consistência e legibilidade.
- Contribuição 3: uma estrutura planejada para ser aplicada em várias áreas, oferecendo uma abordagem genérica e não restrita a casos específicos para lidar com inconsistências nos dados.

Após apresentar as principais contribuições, os autores descrevem as etapas do desenvolvimento do *framework*. A fase inicial é o processamento, que prepara os dados para transformá-los em um formato apropriado para as próximas etapas. Segundo os autores, essa fase de preparação dos dados é essencial para garantir uma correção precisa das irregularidades. Em seguida, os autores detalham as técnicas fundamentais de processamento indispensáveis para as etapas de correção. As etapas de processamento descritas são:

- Seleção de recursos: identificar e selecionar os dados ou atributos mais relevantes, reduzindo a dimensão dos dados e facilitando o processamento.

- Extração de recursos: transformar dados brutos em atributos que capturam informações essenciais, especialmente quando os recursos selecionados não são suficientes para realizar previsões precisas.

- Codificação: converter dados categóricos ou textuais em representações numéricas, permitindo que os modelos preditivos os interpretem adequadamente.

- Normalização e dimensionamento: ajustar os dados para uma escala consistente, facilitando a integração de informações provenientes de diferentes fontes de dados.

- Remoção de espaços em branco e símbolos: limpeza dos dados, eliminando espaços desnecessários e símbolos especiais que não contribuem para a análise, melhorando a consistência e a qualidade geral dos dados.

Após a fase inicial de processamento, inicia-se a fase de seleção de recursos correlacionados, cujo objetivo é identificar e escolher apenas as variáveis ou colunas diretamente relacionadas ao problema de inconsistências nos dados. Quando uma inconsistência é identificada, como uma alteração inesperada no preço de um produto em um conjunto de vendas, é necessário focar apenas nas variáveis ou informações diretamente relacionadas a essa inconsistência (ELOUATAOUI; EL MENDILI; GAHI, 2023).

Em seguida, tem início a fase de seleção da vizinhança apropriada, que consiste em filtrar os registros mais próximos das inconsistências, verificando quais dimensões são afetadas. Para essa etapa, os autores utilizaram as dimensões mencionadas nas contribuições, que são as seis dimensões de qualidade dos dados:

- Completude: refere-se ao grau em que o conjunto de dados contém todas as informações necessárias e esperadas.

- Exatidão: a confiabilidade e conformidade dos dados com os valores esperados ou conhecidos.

- Unicidade e consistência: a unicidade trata dos dados duplicados no conjunto, ou seja, dados que representam a mesma entidade mais de uma vez. A consistência refere-se a informações conflitantes ou contraditórias entre registros duplicados, quando dois registros que se referem à mesma entidade contêm informações diferentes.

- Conformidade: formatos predefinidos, padrões ou especificações definidas para os dados em um conjunto.

- Legibilidade: clareza e compreensão dos dados, garantindo que sejam de fácil entendimento e leitura.

- Previsão de valor correto: após identificar e selecionar os registros relevantes para cada inconsistência, esses registros são atribuídos ao modelo preditivo para realizar uma previsão de valor precisa, preenchendo os dados ausentes e corrigindo valores errados.

Após a etapa de identificação das dimensões afetadas, inicia-se a fase de construção e treinamento do modelo. Para cada tipo de irregularidade identificada, é criado um conjunto de treinamento a partir dos dados válidos (ou seja, aqueles que não apresentam inconsistências). O modelo *XGBoost* é treinado com esses dados para aprender as correlações e padrões necessários para corrigir as irregularidades. O treinamento é realizado para garantir que o modelo aplique as correções de maneira precisa e eficiente. Segundo ELOUATAOUI; EL MENDILI; GAHI (2023), uma das vantagens de usar o modelo *XGBoost* nesta estrutura é sua alta precisão em tarefas de regressão e classificação, tornando-o confiável e adequado para lidar com dados em larga escala.

Os autores aplicaram o *framework* em dois conjuntos de dados. O primeiro é um conjunto sintético (dados gerados artificialmente por algoritmos para substituir os dados reais), com 2 milhões de registros, simulando informações pessoais. O segundo conjunto de dados é o *Titanic*, que contém informações sobre os passageiros do transatlântico britânico *Titanic*, que naufragou no oceano Atlântico em 1912, amplamente utilizado para análise e aprendizado de máquina, e possui 1.309 registros. No artigo, os autores detalham a arquitetura utilizada para implementar o *framework* e as ferramentas utilizadas, na página explicando o funcionamento e o uso dessas ferramentas.

Para avaliar o desempenho da proposta do *framework*, os autores utilizaram duas métricas: exatidão (proporção de correções corretas) e taxa de erro (proporção de correções incorretas). No primeiro conjunto de dados, com consistências predefinidas, as correções foram comparadas diretamente aos valores corretos já conhecidos no conjunto. No segundo conjunto de dados, sem consistências predefinidas, os valores considerados inconsistentes foram definidos manualmente por meio de inspeção cuidadosa, com base no conhecimento do domínio para identificar as anomalias. Os resultados das métricas de exatidão e taxa de erro são apresentados pelos autores no artigo.

Após aplicar o *framework*, os autores discutem os resultados dos testes. O *framework* desenvolvido apresentou um bom desempenho na correção de irregularidades de qualidade em dois conjuntos de dados, obtendo uma exatidão média de 92,71% no primeiro conjunto e 89,45% no segundo. A menor exatidão no segundo conjunto de dados é explicada pelo fato de que conjuntos reais contêm inconsistências mais complexas e diversas. O *framework* mostrou eficácia em dimensões como exclusividade/unicidade, consistência e legibilidade, enquanto conformidade e exatidão foram mais desafiadoras devido à complexidade dos relacionamentos entre elementos de dados. O modelo conseguiu preencher valores ausentes e corrigir valores não conformes, como nomes de classes e gêneros, além de consolidar duplicatas. A melhoria global na qualidade dos dados foi de 18,98%, atingindo uma pontuação de qualidade de 98,5% no segundo conjunto de dados. O *framework* também demonstrou boa escalabilidade, processando grandes volumes de dados com complexidade linear $\mathcal{O}(n)$. Foram identificadas algumas limitações:

- Dificuldade em corrigir valores não conformes sem relação semântica.
- Possibilidade de escolher valores incorretos devido à prevalência.
- Falhas em corrigir anomalias singulares devido à falta de dados comparáveis.
- Aproximação de valores contínuos em vez de predições exatas.
- Incompatibilidade com correções em tempo real, por ser projetado para processar dados já armazenados.

Para os autores, o *framework* se destaca por sua abordagem abrangente e flexível, sendo aplicável a diversas situações e proporcionando uma melhoria na qualidade de grandes volumes

de dados. Mesmo com as limitações, ele oferece uma solução eficaz para corrigir várias inconsistências na qualidade dos dados.

2.2.3 Big Data Quality: A Quality Dimensions Evaluation

A maioria das grandes e pequenas empresas considera que quase todas as decisões estratégicas de negócios são baseadas em informações extraídas dos dados (TALEB et al., 2016). Para os autores TALEB et al. (2016), os dados na sua forma original, sem nenhum tratamento prévio, podem estar incompletos e ter muitas inconsistências, que podem ser causadas por alguns fatores, incluindo o fator humano. Em ambientes que lidam com grande volume de dados, os dados são elementos importantes por passarem pelo processamento de dados e de extração de informações. Para os autores TALEB et al. (2016), dados inconsistentes ou inapropriados podem gerar analises tendenciosas, causada por má preparação dos dados, natureza dos dados, incluindo formato e tipo.

Dessa forma, é importante analisar a qualidade dos dados, para identificar inconsistências nos dados e aplicar melhorias. Os autores propõem um esquema rápido de avaliação da qualidade de *Big Data*, aplicando uma estratégia de amostragem em um grande conjunto de dados, usando uma amostragem representativa, para uma rápida análise da qualidade dos dados. Os autores explicam que avaliar a qualidade dos dados de *Big Data* é justificada pelo impacto que dados com irregularidades afetam os resultados analíticos, esses resultados analíticos são a extração de informação para tomadas de decisões. Para os autores existem elementos importantes para lidar com a avaliação e melhoria da qualidade de dados, sendo conhecidos e categorizados sob Dimensões de qualidade dos dados (DQD), as comuns mencionadas pelos autores são:

- Exatidão: o qual próximo o dado coletado está conforme a informação real, dados sem erros no cadastro.
- Completude: a presença de todos os valores necessários e esperados em um conjunto de dados, informações estão completas.
- Consistência: regras semânticas definidas sobre os dados, com referência a teoria relacional, sendo suas regras de restrições no cadastro da informação.

Para medir algumas dessas DQD, utiliza-se uma estratégia orientada a dados. As medições são aplicadas diretamente sobre os próprios dados, e cada métrica é projetada especificamente para avaliar uma determinada DQD. Os autores apresentaram métricas correspondentes a cada dimensão da qualidade dos dados, o resultado dessas métricas é expresso em porcentagem.

Após definirem os elementos necessários para analisar a qualidade dos dados e identificar as dimensões afetadas, os autores apresentaram a proposta de um esquema para essa análise, bem como sua aplicação prática.

Os autores apresentada como o esquema funciona. Como mencionado no início, a proposta é criar um esquema rápido de avaliação da qualidade de dados, antes de realizar analises nos dados, o esquema funciona considerando a qualidade das DQD aplicando as métricas para medir a qualidade, as DQD é exatidão, integridade ou/e consistências. Para aplicação do esquema os autores utilizaram um conjunto de dados chamado *Sleep Heart Health Study* um conjunto para avaliar os efeitos da respiração desordenada do sono, este conjunto de dados é coletado de 6.441 pessoas, contendo atributos como; eletrocardiograma, eletroencefalograma, eletrooculograma, eletromiografia, excursões torácicas e abdominais, fluxo de ar nasal, saturação de oxigênio, eletrocardiograma e frequência cardíaca, totalizando 1.278 atributos, com arquivos em Formato de Dados Europeu (EDF) de 40 Capacidade de armazenamento de dados em

sistemas computacionais (MB) por paciente. Os autores propõem um esquema de avaliação da qualidade de dados baseado em *Bag of Little Bootstraps (BLB)*, que reduz o tempo e os recursos computacionais, gerando ações para melhorar a qualidade dos dados. os autores apresentam o algoritmo desenvolvido para avaliar a qualidade dos dados em ambiente de *Big Data*.

Dois cenários foram analisados:

- Primeiro cenário: verifica a completude, analisando a presença de valores ausentes, e revela que 80% dos atributos têm menos de 60% de dados faltantes. As ações sugeridas incluem descartar linhas ou colunas com 80% ou mais de dados ausentes e substituir valores faltantes pela média.

- Segundo cenário: avalia a consistência, verificando a conformidade dos dados com restrições. Apenas 5% dos atributos têm mais de 90% de dados ausentes, enquanto 29,1% não possuem valores ausentes, garantindo 100% de consistência se mantidos isoladamente. No entanto, a consistência geral usando todos os atributos é de apenas 29,1%.

Como trabalho futuro, os autores propõem o desenvolvimento de algoritmos automáticos para gerar e otimizar métricas de qualidade dos dados.

2.2.4 Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context

A *Big Data* tem quatro aspectos: volume, velocidade, variedade e veracidade. A veracidade é uma característica do *Big Data* que ganhou popularidade e se refere à qualidade dos dados envolvidos (JUDDOO et al., 2018). O setor de saúde é um exemplo de indústria que utiliza dados em grande quantidade, os quais podem ser categorizados como registros eletrônicos de saúde, dados administrativos, dados de reivindicações, entre outros. O *Big Data* está sendo aplicado para melhorar a tomada de decisões no setor de saúde (JUDDOO et al., 2018).

Os autores JUDDOO et al. (2018) citam o *Evidence-based medicine (EBM)*, que consiste no uso de dados provenientes da medicina baseada em evidências, utilizando as melhores evidências disponíveis para apoiar decisões no cuidado de pacientes. O Suporte à Decisão Personalizados (SSDP), que está aprimorando essa prática por meio da análise de *Big Data*. Os autores também mencionam que, devido à variedade de grandes conjuntos de dados médicos, as empresas farmacêuticas têm recorrido à análise de *Big Data* para descobrir novos medicamentos e compreender certas doenças. Existe a utilização da *Internet of Things (IoT)* na área da saúde, com dados coletados por meio da *IoT*, integrando *Big Data* e *IoT* (JUDDOO et al., 2018).

Porém, os autores JUDDOO et al. (2018) ressaltam que, sem a qualidade adequada dos dados, o uso de *Big Data* no setor da saúde não será eficaz. Entender as dimensões da qualidade dos dados é essencial para mensurar a qualidade e implementar qualquer iniciativa voltada para sua melhoria. No artigo *Data Governance in the Health Industry: Investigating Data Quality Dimensions within a Big Data Context*, os autores têm como principais objetivos criar uma discussão e recomendar DQD mais relevantes e adequadas no contexto de *Big Data* no setor da saúde, além de validar a importância de cada dimensão identificada. Após a análise de artigos relacionados ao tema, os autores apresentam algumas DQD específicas para o setor da saúde. Para isso, eles realizaram a leitura de 41 artigos, atribuindo a cada um peso conforme a relevância das DQD no contexto de *Big Data*.

- Peso baixo (L, valor 1): atribuído quando nenhum dos contextos (*Big Data* ou saúde) estava presente, mas havia alguma conexão entre as DQD e o setor.

- Peso médio (M, valor 2): atribuído quando apenas um dos contextos (*Big Data* ou saúde) estava presente.
- Peso alto (H, valor 3): atribuído quando ambos os contextos (*Big Data* e saúde) estavam claramente presentes.

Após a análise dos artigos, os autores identificaram que a exatidão é uma das dimensões mais frequentemente citadas. A contagem ponderada revelou as dimensões mais significativas, confirmando essa tendência. Conforme JUDDOO et al. (2018), é apresentada uma hierarquia das dimensões mais discutidas, classificadas de acordo com seu peso e relevância nos contextos de *Big Data* e saúde. Essa hierarquia foi organizada em formato de tabela, conforme ilustrado no quadro 2.3, que apresenta uma adaptação da tabela original.

Quadro 2.3: Contagens das diferentes dimensões de qualidade de dados (DQD)

DQD	**Contagem ponderada total**	**DQD**	**Contagem ponderada total**
Exatidão	58	Legibilidade	2
Completude	52	Captura de dados	2
Consistência	30	Privacidade	2
Confiabilidade	15	Heterogeneidade	2
Pontualidade	11	Proveniência	2
Atualidade	8	Abrangência	2
Disponibilidade	6	Concordância	2
Acessibilidade	5	Confiança	2
Confiabilidade (sentido de confiança)	5	Volume de dados	2
Segurança	5	Coincidência	2
Corretude	5	Verificabilidade	2
Plausibilidade	4	Compreensibilidade	2
Relevância	4	Rastreabilidade	2
Clareza	4	Conformidade	2
Validade	4	Cobertura	2
Exclusividade/unicidade	3	Usabilidade	2
Formato	3	Qualidade da apresentação	2
Precisão	3	Expansibilidade	2
Utilidade	3	Sensibilidade	2

Fonte: Adaptado de RIBEIRO et al, 2021.

Conforme apresentado no quadro 2.3, a completude e a exatidão foram as dimensões mais citadas nos artigos analisados pelos autores. Após essa análise, JUDDOO et al. (2018) destacam que as DQD mais relevantes são: exatidão, completude, consistência, confiabilidade e pontualidade. Os resultados foram avaliados com base no *framework* de Wang e Strong. WANG; STRONG (1996), que organiza as dimensões em quatro categorias de qualidade de dados: intrínseca, contextual, representacional e acessibilidade. As contagens totais das DQD em cada categoria foram apresentadas no artigo em formato de tabela. O quadro 2.4 mostra uma versão adaptada dessa tabela.

Os autores investigaram as DQD mais relevantes no contexto de *Big Data* na indústria da saúde. A pesquisa foi conduzida com base no ciclo *Inner Hermeneutic Cycle (IHC)*, abordando quatro questões principais. As questões de pesquisa foram:

Quadro 2.4: Dimensões de categoria DQD com agregados de contagem.

Categoria	Dimensões Individuais	Contagem
Intrínseca	Exatidão, Confiança, Plausibilidade, Precisão, Conformidade, Rastreabilidade, Verificabilidade, Proveniência, Confiança, Concordância, Correção	87
Contextual	Completude, Pontualidade, Atualidade, Confiabilidade, Disponibilidade, Exclusividade, Relevância, Validade, Expansibilidade, Sensibilidade, Cobertura, Volume de Dados, Abrangência, Heterogeneidade	108
Representacional	Consistência, Formato, Utilidade, Legibilidade, Captura, Coincidência, Compreensibilidade, Usabilidade, Qualidade de Apresentação	48
Acessibilidade	Acessibilidade, Segurança, Privacidade, Conformidade	14

Fonte: Adaptado de RIBEIRO et al, 2021.

- RQ1: a categoria intrínseca de DQD, que inclui exatidão e confiabilidade, é aplicável em todas as situações em que os dados são utilizados, incluindo *Big Data* e o setor da saúde. No entanto, sua importância não é tão alta quanto a de outras categorias. Isso sugere que a qualidade dos dados não deve ser considerada de forma isolada, pois depende das aplicações de *software* e dos usuários envolvidos.
- RQ2: a categoria contextual é a mais importante, pois, devido à variedade de dados e consumidores no setor da saúde, exige diferentes critérios de qualidade para cada uso de dados. O contexto, portanto, é essencial para aplicações que envolvem qualidade de dados.
- RQ3: a categoria representacional tem menor importância em relação às duas categorias anteriores, sendo mais relevante para dados baseados em texto e números. A diversidade de tipos de dados, como imagens e vídeos, precisa ser investigada mais a fundo.
- RQ4: a categoria acessibilidade apresenta a menor importância, uma vez que muitos dados estão publicamente disponíveis. Contudo, o grande volume de dados pode impactar o acesso, especialmente em contextos privados, onde segurança e privacidade são fatores críticos.

Após todos os estudos e análises dos artigos, os autores JUDDOO et al. (2018) concluem que a pesquisa foi conduzida em um contexto multidisciplinar, envolvendo três campos principais: QD, GD, *Big Data* e informática em saúde. *Big Data* e informática em saúde são dois campos que, segundo os autores, têm poucas pesquisas com foco na perspectiva de qualidade de dados, o que eles consideram importante, pois a qualidade gera mais valor aos dados. O objetivo dos autores com a pesquisa discutida no artigo foi investigar quais DQD podem ser mais importantes no contexto de *Big Data* no setor da saúde, utilizando o *IHC* adotado como método de pesquisa. Os autores obtiveram como resultado que as DQD mais relevantes são: exatidão, completude e consistência, e identificaram a confiabilidade e pontualidade como importantes. A categoria contextual foi considerada a mais relevante, explicada pela variedade de dados e aplicações de *Big Data* na saúde.

O estudo é um dos primeiros a usar um método sistemático para identificar as DQD mais importantes no contexto de *Big Data* na saúde. Os resultados podem ser validados em diferentes contextos, já que a qualidade dos dados é altamente contextual. A limitação do volume de literatura é uma das principais limitações do estudo.

A pesquisa serve de base para futuras investigações, como o uso de DQD para algoritmos de aprendizado de máquina que distinguem dados de qualidade de dados não qualificados. Métodos estatísticos mais aprimorados, como a análise semântica *Latent Semantic Analysis (LSA)*, para determinar as dimensões DQD mais importantes, serão usados em futuras pesquisas. Pesquisas empíricas também serão realizadas com uma maior variedade de dados para explorar as categorias de QD representacional e acessibilidade, e as cinco principais dimensões de DQD serão usadas para melhorar algoritmos de aprendizado de máquina e desenvolver um algoritmo de aprendizagem de máquina mais eficiente na classificação de dados incorretos e no reparo de dados.

3
Referencial Teórico

Neste capítulo são abordados conceitos essenciais para o desenvolvimento deste trabalho, incluindo as metodologias utilizadas, seu funcionamento, importância e aplicação. Essa seção fornecerá uma base teórica de apoio para o desenvolvimento do trabalho.

3.1 Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec)

O Sistec é um sistema do governo brasileiro, instituído e implantado pelo MEC em 2009, por intermédio da Secretaria de Educação Profissional e Tecnológica do MEC (Ministério da Educação, 2024b). O objetivo do Sistec é operar como ferramenta para registro e divulgação dos dados da Educação Profissional e Tecnológica, além de validar diplomas de cursos de educação profissional técnica de nível médio. O Sistec também apoia o planejamento da oferta de cursos gratuitos viabilizados pela, Secretaria de Educação Profissional e Tecnológica - MEC (Setec-MEC), por meio da bolsa formação do Pronatec (Ministério da Educação, 2024b).

Ele armazena dados de instituições públicas e privadas e de matrículas, servindo também como auxílio para o planejamento da oferta de cursos, matrículas, confirmação de frequência e de concluintes. As instituições de ensino ofertantes de EPT inserem as informações sobre os cursos técnicos de nível médio e os cursos de qualificação profissional, incluindo matrícula, frequência, concluintes, entre outros dados (Ministério da Educação, 2024b).

O preenchimento dos dados é obrigatório e exerce papel fundamental para garantir a validade nacional dos diplomas expedidos. Esse preenchimento é obrigatório para todas as unidades de ensino credenciadas para oferta de cursos de, EPT, independentemente de sua dependência administrativa (ou seja, pública ou privada).

3.2 Big Data

Big Data é um termo utilizado para se referir ao armazenamento de dados em grande volume, com alta variedade de informações e complexidade. Segundo ULARU et al. (2012), o conceito de *Big Data* foi introduzido pela primeira vez no mundo da computação aproximadamente em 2005, por Roger Magoulas, que o definiu como um conjunto de dados caracterizado por grande volume e alta complexidade. Os autores também mencionam a visão da *Big Data* sobre os aspectos do *Big Data*.

A *Big Data*, uma das primeiras empresas de tecnologia, existente desde 1896, destaca os seguintes aspectos fundamentais do *Big Data*: volume, velocidade, variedade e veracidade (ULARU et al., 2012). Esses aspectos são detalhados no quadro 3.1.

Para ULARU et al. (2012), a principal importância do *Big Data* está na sua capacidade de melhorar a eficiência no uso de grandes volumes de dados de diferentes tipos. O *Big Data* oferece melhor desempenho no armazenamento e processamento de dados, permitindo que as informações coletadas sejam utilizadas de forma mais eficiente para a extração de informações

Quadro 3.1: Aspectos da Big Data

Nome	Descrição
Volume	A quantidade de dados armazenados para obter algum conhecimento.
Velocidade	Tempo do qual a *Big Data* pode processar e realizar alguma atividade, principalmente em atividades que necessitam de rápidas respostas.
Variedade	Tipo dos dados armazenados, que podem ser dados estruturados e não estruturados ou semi estruturados.
Veracidade	Grau em que as informações são confiáveis para serem utilizadas para tomar decisões, utilizar dos dados para extrair informações relevantes.

Fonte: Adaptado de ULARU et al, 2012.

para tomadas de decisões. Seus aspectos auxiliam no gerenciamento de dados com alto volume e variedade.

Embora todos os aspectos do *Big Data* sejam importantes, o de variedade se destaca como um dos mais relevantes no gerenciamento de dados. Esse aspecto engloba dados estruturados, semi-estruturados e não estruturados:

- Dados estruturados: organizam-se em um formato altamente específico e pré-definido.
- Dados semi-estruturados: não seguem um padrão rígido, mas possuem alguma organização que facilita sua coleta e armazenamento.
- Dados não estruturados: não possuem um padrão definido, tornando-se mais desafiadores para lidar.

Dessa forma, o *Big Data* organiza esses dados, facilitando sua coleta e processamento, garantindo maior velocidade e veracidade das informações. Com o suporte de tecnologias avançadas, é possível processar e extrair valor desses dados de maneira mais eficiente.

3.3 Governança de Dados

Toda empresa, seja pública ou privada, que utiliza dados para a tomada de decisões necessita da criação de regras e políticas para a gestão desses dados, incluindo normas de coleta, acesso e alteração. A GD estabelece diretrizes para uma gestão eficiente dos dados. Segundo PANIAN (2010), a definição de GD abrange processos, políticas, padrões organizacionais e tecnologias que auxiliam na administração, garantindo a disponibilidade, acessibilidade, qualidade, consistência, auditabilidade e segurança dos dados em uma organização.

Essas políticas têm como responsabilidades o gerenciamento, que apoia a tomada de decisões, a segurança dos dados, por meio da criação de normas de acesso e alteração, a padronização de processos a serem seguidos por toda a empresa e a melhoria da eficácia no uso dos dados, especialmente por meio da redução de custos e da prevenção de falhas.

Embora cada empresa deva definir suas próprias políticas e necessidades para gerenciar seus dados, PANIAN (2010) apontam que, para uma boa GD, existem quatro componentes importantes a serem seguidos:

- Padrões: estabelecer os padrões que os dados da empresa devem seguir.
- Políticas e processos: definir e cumprir as políticas de processos e gerenciamento, estabelecendo regras para os dados e auditoria. Também é essencial implementar mecanismos de monitoramento e medição, além de gerenciar alterações e acessos aos dados.
- Organização: estruturar responsabilidades para cada área, definindo claramente o papel de cada setor na aplicação das políticas de dados e na gestão desses dados.
- Tecnologia: adotar tecnologias adequadas para a captura e gerenciamento dos dados, buscando ferramentas que agilizem o trabalho e sejam eficazes.

Esses componentes servem como suporte, sendo aplicados conforme a necessidade dos dados e da empresa. Eles permitem aprimorar o gerenciamento, a segurança e o processamento dos dados. O principal objetivo da GD é garantir que as informações em uma organização sejam gerenciadas de forma eficaz, segura e conforme as políticas estabelecidas.

3.4 Qualidade dos Dados

Segundo ELOUATAOUI et al. (2022); CAI; ZHU (2015), a pesquisa sobre qualidade dos dados começou em 1990, com algumas definições diferentes. Eles mencionam que um grupo chamado *total data quality management*, da *Massachusetts Institute of Technology (MIT) University*, liderado pelo Professor Richard Y. Wang, definiu a qualidade dos dados como "adequação para uso". Eles também definiram as dimensões da qualidade dos dados como um conjunto de atributos que representam um único aspecto ou construção da qualidade dos dados.

Medir a qualidade dos dados não é fácil, especialmente no contexto de *Big Data*, onde o volume, a variedade e a velocidade dos dados tornam o processo ainda mais desafiador. As DQD funcionam como um padrão amplamente reconhecido e aceito para avaliar a qualidade dos dados, fornecendo critérios claros e mensuráveis que permitem analisar a qualidade das informações e determinar se os dados são adequados para o uso. Para avaliar a qualidade dos dados, utilizam-se as dimensões e as métricas associadas a essas dimensões. As dimensões escolhidas dependem das necessidades dos dados e de suas características. Após definir as dimensões, utilizam-se as métricas para mensurar a qualidade de cada dimensão.

Os autores ELOUATAOUI et al. (2022) apresentam cinco Avaliação da Qualidade dos Dados (DQA). O quadro 3.2 apresenta essas cinco mencionadas pelos autores em formato de quadro. Também é apresentado DQD e as doze métricas para medir as dimensões mencionadas pelos autores, no quadro 3.3.

Quadro 3.2: Aspectos da Qualidade dos Dados

DQA	Descrição
Confiabilidade	Confiabilidade e credibilidade dos dados.
Disponibilidade	Acessibilidade e compartilhamento dos dados, mantendo o nível adequado de segurança.
Usabilidade	Relevância e à facilidade de uso dos dados.
Validade	Assegura que os dados estão em um formato específico e cumprem as regras de negócios definidas.
Pertinência	Refere-se à adequação e adequação dos dados ao contexto de uso.

Fonte: Adaptado de RIBEIRO et al, 2021.

Quadro 3.3: Métricas de qualidade de dados (DQA e DQD)

DQA	DQD	Descrição DQD	Métricas
Confiabilidade	**Integridade**	Mede a precisão dos dados, comparando dados originais e processados.	$\frac{\text{Diferenças entre valores originais e processados}}{\text{Total de valores}} \times 100$
	Volatilidade	Mede o tempo que os dados permanecem válidos.	$\frac{\text{Data de Criação - Data de Modificação}}{\text{Data Atual - Data de Criação}} \times 100$
Disponibilidade	**Acessibilidade**	Mede a qualidade de estar disponível e de fácil pesquisa.	$\frac{\text{Valores acessíveis}}{\text{Total de valores}} \times 100$
	Segurança	Mede a proteção dos dados, considerando políticas de acesso e criptografia.	-
Usabilidade	**Completude**	Mede a extensão dos dados completos e atendendo necessidades.	$\frac{\text{Número de valores não vazios}}{\text{Total de valores}} \times 100$
	Relevância	Mede a medida da relevância dos dados para as análises.	$\frac{\text{Número de acessos ao campo}}{\text{Total de acessos à tabela que inclui campo}} \times 100$
	Facilidade de Manipulação	Mede a facilidade de trabalhar com dados após o pré-processamento.	$\frac{\text{Número de diferenças entre a tabela original e a limpa}}{\text{Total de dados}} \times 100$
Validade	**Consistência**	Mede a coerência dos dados com tipos e esquemas definidos.	$\frac{\text{Número de valores com tipos consistentes}}{\text{Total de valores}} \times 100$
	Legibilidade	Mede a capacidade de processar os dados sem erros semânticos.	$\frac{\text{Número de valores processados e sem erros de digitação}}{\text{Total de valores}} \times 100$
	Conformidade	Mede a aderência às regras de formato (ex. datas, telefones).	$\frac{\text{Número de valores com formato consistente}}{\text{Total de valores}} \times 100$
Pertinência	**Unicidade**	Mede a redundância nos dados.	$\frac{\text{Número de linhas únicas}}{\text{Total de valores}} \times 100$
	Pontualidade	Mede a atualização dos dados, medindo o tempo desde a última modificação.	$\frac{\text{Data Atual - Ultima Data de Modificação}}{\text{Data Atual - Data de Criação}} \times 100$

Fonte: Adaptado de RIBEIRO et al, 2021.

De acordo com ELOUATAOUI et al. (2022, p. 9), "diferentes critérios podem ser usados para classificar e agrupar as métricas de qualidade, como a natureza das métricas, o significado das métricas e até mesmo o contexto do estudo". Com a definição das DQA e DQD, é possível utilizá-las para categorizar os dados e aplicar métricas para medir sua qualidade. Essas DQA e DQD ajudam a entender quais dimensões dos dados foram afetadas por inconsistências, além de fornecer métricas que avaliam, em porcentagem, o impacto dessas inconsistências na qualidade dos dados.

Após medir a qualidade dos dados, é possível adotar melhorias, implementando ações corretivas para reduzir as inconsistências e, assim, aprimorar a qualidade dos dados. Ter uma boa qualidade de dados é fundamental para extrair informações consistentes e precisas para a tomada de decisões baseadas em dados.

A segurança dos dados é medida por alguns critérios que são;

- Política de acesso: restringe o uso dos dados, garantindo que apenas pessoas autorizadas possam acessá-los (20%).
- Protocolos de segurança: uso de protocolos para garantir a transferência segura de dados, como criptografia de dados durante a transmissão (20%).
- Detecção de ameaças: medidas para identificar potenciais ameaças ou acessos não autorizados aos dados (20%).
- Criptografia: protege os dados por meio de criptografia, garantindo que estejam protegidos contra acessos não autorizados (20%).
- Documentação de segurança: disponibilidade de documentação que especifique as políticas de segurança e como os dados devem ser manipulados e protegidos (20%).

4
Metodologia

Neste capítulo, será abordada a metodologia utilizada para o desenvolvimento deste trabalho e ferramentas e tecnologias que servira de apoio. Será apresentada a fonte de dados utilizada, bem como a metodologia aplicada tanto para a análise quanto para a sugestão de melhorias na qualidade dos dados.

4.1 Ferramentas e tecnologias

Nesta seção, são apresentadas as principais tecnologias e ferramentas utilizadas para o desenvolvimento deste trabalho.

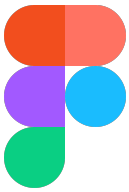
4.1.1 Ferramentas para desenvolvimento das documentações

As principais ferramentas utilizadas para a criação das documentações das regras de negócio foram:

- Figma: ferramenta de design colaborativo para criação e compartilhamento de protótipos e diagramas visuais.

O **Figma** foi escolhido por permitir o desenvolvimento de protótipos e diagramas de forma visual e interativa, facilitando o trabalho colaborativo entre os membros da equipe (FIGMA, 2025).

Figura 4.1: Logo do Figma, ferramenta de design colaborativo.



Fonte: https://www.figma.com/community/file/1252952214981489076

4.1.2 Para o desenvolvimento dos scripts

Na área da programação, *scripts* são conjuntos de instruções escritas em uma linguagem de programação, com o objetivo de automatizar tarefas executadas por um computador. Para o desenvolvimento dos *scripts* utilizados na validação dos dados, com base nas respectivas regras de negócio, foi utilizada a linguagem Python. É uma linguagem de programação de alto nível, com sintaxe simples e clara, que facilita a criação de *scripts* voltados para automação de tarefas, manipulação de dados e integração entre sistemas (PYTHON DOCUMENTATION, 2025).

Figura 4.2: Logo da linguagem Python, utilizada para automação de tarefas.

Fonte: https://www.python.org/community/logos/

4.1.3 Para geração de visualizações e tabelas

Para a criação de visualizações gráficas e tabelas para o resultados das análises das inconsistências, foram utilizadas algumas bibliotecas, que são coleções de código que podem ser importadas e usadas em diferentes projetos. As bibliotecas utilizadas foram `pandas`, `matplotlib` e `seaborn`.

A biblioteca `pandas` é amplamente utilizada para manipulação e análise de dados, permitindo o carregamento de arquivos em diversos formatos, além de filtragem e agregação das informações. A `matplotlib` é voltada para a criação de gráficos, possibilitando a geração de visualizações personalizadas. Já a `seaborn`, construída sobre a `matplotlib`, facilita a criação de gráficos estatísticos com uma melhor aparência e gráficos intuitivos.

Figura 4.3: Logos das bibliotecas `pandas`, `matplotlib` e `seaborn` utilizadas para manipulação e visualização de dados.

Fonte: https://pandas.pydata.org/about/citing.html, https://matplotlib.org/stable/gallery/misc/logos2.html, https://seaborn.pydata.org/citing.html

4.2 Fonte de dados

4.2.1 Dados disponíveis

Os dados utilizados para a análise da qualidade foram os registros de matrículas do Sistec IFB. Esses dados estão disponíveis ao público em formato Comma-separated values (CSV), no portal de diretórios do IFB, acompanhados de um dicionário que descreve o significado de cada variável. Esse dicionário é apresentado no quadro 4.5 e contém o nome das colunas, bem como a descrição do que cada valor representa. O conjunto de dados possui 120.873 registros, contendo matrículas desde o ano de 2009 até 2025, com 40 colunas. Sua última atualização ocorreu em 25 de março, às 21:09, no horário de Brasília.

4.2.2 Solicitação de dados adicionais

As solicitações de dados adicionais foram realizadas por meio da plataforma Fala.BR, "uma plataforma integrada de ouvidoria e acesso à informação do Poder Executivo Federal, que permite o envio de pedidos de acesso à informação e manifestações de ouvidoria (denúncias, elogios, reclamações, sugestões e solicitações) aos órgãos e entidades"(BRASIL, 2025).

Além dos dados disponíveis em arquivos CSV descritos na seção anterior, foram realizadas oito solicitações de dados junto a outras instituições por meio da plataforma Fala.BR. Todas as solicitações receberam resposta, porém não foi possível obter os dados solicitados devido a restrições relacionadas à extração ou ao tratamento de informações sensíveis, bem como à ausência de colunas necessárias para análise nos dados disponibilizados por algumas instituições. Dessa forma, esta análise baseia-se exclusivamente nos dados descritos na seção anterior 4.2.1, que são os dados provenientes do Sistec do IFB.

4.3 Fases para analisar as inconsistências relatadas

Para analisar os dados do Sistec referentes às matrículas do IFB, foram aplicadas três fases principais, compostas por etapas que auxiliam na análise e na medição da qualidade dos dados. Conforme explicado por BATINI et al. (2009), existem diversas perspectivas para analisar e comparar a qualidade dos dados, mas, de forma geral, essas três fases são as principais:

1. Reconstrução do estado: coleta informações contextuais sobre processos, gestão e qualidade.

2. Avaliação/medição: mede a qualidade dos dados e compara com referências para diagnosticar problemas. De acordo com BATINI et al. (2009), essa fase inclui as seguintes etapas:

    - Análise de dados: examina esquemas de dados e realiza entrevistas para alcançar. uma compreensão completa dos dados e das regras dos dados e de gerenciamento relacionadas.
    - Análise de requisitos: pesquisa a opinião dos usuários e administradores responsáveis pelos dados, para identificar questões de qualidade e definir novas metas de qualidade
    - Identificação de áreas críticas: que seleciona as bases de dados, dados mais relevantes, fluxos a serem avaliados quantitativamente.
    - Modelagem de processos: que fornece um modelo dos processos que produzem ou atualizam dados.
    - Medição da qualidade: define métricas quantitativas ou qualitativas para avaliar dimensões de qualidade.pode ser objetiva quando baseada em métricas quantitativas, ou subjetiva, quando é baseada em avaliações qualitativas feitas por administradores de dados e usuários

3. Melhoria: seleciona estratégias e técnicas para atingir novas metas de qualidade.

Para analisar as inconsistências relatadas nos dados do Sistec IFB e medir a qualidade dos dados, foram aplicadas as fases mencionadas anteriormente. Conforme apresentado pelos autores BATINI et al. (2009), diversas metodologias contemplam etapas que auxiliam nas fases de reconstrução, avaliação e melhoria da qualidade dos dados.

Na etapa de melhoria, algumas metodologias contêm apenas ações baseadas em dados, outras são focadas apenas em processos, e algumas são mistas ás quais são as mais completas. Mas, cada metodologia é indicada para um tipo específico de objetivo ao analisar e melhorar a qualidade dos dados. O quadro 4.1 apresenta essas metodologias e o que cada uma cobre nas fases de avaliação da qualidade dos dados.

Quadro 4.1: Comparação das metodologias para avaliação da qualidade dos dados

Método	Análise de dados	Análise de requisitos de DQ	Identificação de áreas críticas	Modelagem de processos	Medição de qualidade	Extensível a outras dimensões e métricas
TDQM	+		+	+	+	Fixo
DWQ	+	+	+		+	Aberto
TIQM	+	+	+	+	+	Fixo
AIMQ	+		+		+	Fixo
CIHI	+		+			Fixo
DQA	+		+		+	Aberto
IQM	+				+	Aberto
ISTAT	+				+	Fixo
AMEQ	+		+	+	+	Aberto
COLDQ	+	+	+	+	+	Fixo
DaQuinCIS	+		+	+	+	Aberto
QAFD	+	+	+		+	Fixo
CDQ	+	+	+	+	+	Aberto

Fonte: BATINI et al, 2009

A seguir, são apresentados os principais métodos utilizados para avaliação e gerenciamento da qualidade dos dados. Cada um desses métodos contribui com abordagens específicas para identificar, medir e melhorar diferentes aspectos da qualidade da informação, conforme será detalhado ao longo do trabalho.

- Gestão Total da Qualidade dos Dados (TDQM)
- Metodologia da Qualidade de Data Warehouse (DWQ)
- Gestão Total da Qualidade da Informação (TIQM)
- Metodologia para Avaliação da Qualidade da Informação (AIMQ)
- Metodologia do Instituto Canadense de Informação em Saúde (CIHI)
- DQA
- Medição da Qualidade da Informação (IQM)
- Metodologia ISTAT (ISTAT)
- Metodologia Baseada em Atividades para Medição e Avaliação da Qualidade da Informação de Produto (AMEQ)
- Metodologia Loshin (Custo do Baixo Nível de Qualidade dos Dados) (COLDQ)
- Qualidade dos Dados em Sistemas Cooperativos de Informação (DaQuinCIS)
- Metodologia para Avaliação da Qualidade de Dados Financeiros (QAFD)
- Metodologia Abrangente para Gestão da Qualidade dos Dados (CDQ)

Compreender o que essas metodologias cobrem nas fases de avaliação da qualidade dos dados é fundamental para escolher a mais adequada para orientar o processo de análise. BATINI et al. (2009) detalham o que cada metodologia contempla em relação à avaliação e

Quadro 4.2: Etapas de melhoria das metodologias

Método	Controle do processo	Projeto de soluções de melhoria de dados	Redesenho para melhoria do processo	Gerenciamento da melhoria	Monitoramento da melhoria
TDQM			+	+	+
DWQ		+		+	
TIQM		+	+		+
DQA					
ISTAT		+	+		
AMEQ					+
COLDQ	+	+	+		+
DaQuinCIS					
CDQ	+	+	+		

Fonte: BATINI et al, 2009

quais dimensões de qualidade são consideradas sejam elas subjetivas, objetivas ou ambas. O quadro 4.2 apresenta uma comparação das metodologias quanto à cobertura da fase de melhoria da qualidade dos dados, que envolve estratégias orientadas a dados, a processos ou ambas.

Algumas metodologias não contemplam etapas de melhoria, apenas avaliação, o que justifica o número menor de metodologias apresentadas no quadro 4.2. Algumas coberturas são mais completas que outras.

Neste trabalho, foi escolhida e aplicada a metodologia *CDQ*, caracterizada por três fases principais: reconstrução do estado, avaliação e escolha do processo ótimo de melhoria, fases que foram aplicadas neste trabalho como reconstrução, avaliação e melhoria. A escolha da *CDQ* se deu porque essa metodologia abrange, de forma completa, todas as fases, tanto aquelas realizadas durante o projeto de pesquisa quanto as aplicadas neste trabalho, para analisar a qualidade dos dados.

Na fase de melhoria, as ações não foram aplicadas diretamente aos dados do Sistec IFB, mas sim apresentadas como sugestões, baseadas nas inconsistências identificadas e nas dimensões da qualidade afetadas pelas irregularidades. A metodologia *CDQ* foi utilizada como base teórica para as etapas de reconstrução e avaliação, servindo também como referencial para a proposição de recomendações de melhorias futuras. Dessa forma, ela funciona como uma metodologia completa e eficaz, atendendo às necessidades de aprimoramento da qualidade dos dados identificadas nas fases anteriores.

Fase de reconstrução

A primeira fase foi a de reconstrução, realizada durante o projeto de pesquisa mencionado na seção de justificativa e ajustada para utilização neste trabalho, com o objetivo de analisar a qualidade dos dados. No projeto mencionado, foi realizado um estudo detalhado dos dados, com o desenvolvimento da documentação de cada coluna baseada nas regras de negócio. Várias reuniões com o gestor do Sistec foram realizadas ao longo do projeto para atualizar essa documentação, esclarecer dúvidas, revisar as regras de negócio e garantir sua correção.

Como os dados do Sistec seguem uma estrutura padrão para matrículas no Brasil, a documentação desenvolvida mantém um formato que também se aplica aos dados do Sistec IFB. Essa documentação inclui o dicionário de dados, a descrição de cada coluna e as regras

que os dados devem seguir no banco de dados como, por exemplo, o formato esperado, se o valor é numérico ou textual, e se pode ou não ser vazio. Essa documentação é essencial para compreender as regras de negócio associadas aos dados e criar validações que verifiquem se eles estão em conformidade com essas regras.

Fase de avaliação/medição

Para a fase de avaliação contém etapas que auxiliam para aplicação da avaliação/medição que são:

- Análise de dados: como mencionado na etapa de reconstrução, foram realizadas reuniões para entender os dados do Sistec e suas regras de negócio. Essa análise foi realizada no projeto de pesquisa e é aplicada aos dados do Sistec IFB utilizados neste trabalho, já que ambos seguem a mesma estrutura padrão.

- Análise dos requisitos de qualidade dos dados: durante o projeto de pesquisa, entrevistas com o gestor do Sistec resultaram em sugestões e esclarecimento de quais validações nos dados era necessária realizar.

- Identificação de áreas críticas: são selecionadas as colunas e informações que, após a validação com as regras de negócios, apresentaram inconsistências.

- Modelagem de processos: compreender como os dados do Sistec IFB são coletados, produzidos e atualizados para analisar a qualidade dos dados.

- Funções de validação foram desenvolvidas para identificar inconsistências, com base nas regras de negócio definidas anteriormente.

- Irregularidades identificadas são relatadas e categorizadas por dimensões de qualidade afetadas.

- Cada dimensão é avaliada com métricas específicas para medir a qualidade dos dados.

Neste trabalho, foram selecionadas 13 dimensões de qualidade para analisar e medir a qualidade dos dados do Sistec IFB. O quadro 4.3 apresenta as 13 dimensões juntamente com suas respectivas métricas. A métrica que calcula segurança é apresentada na seção qualidade de dados

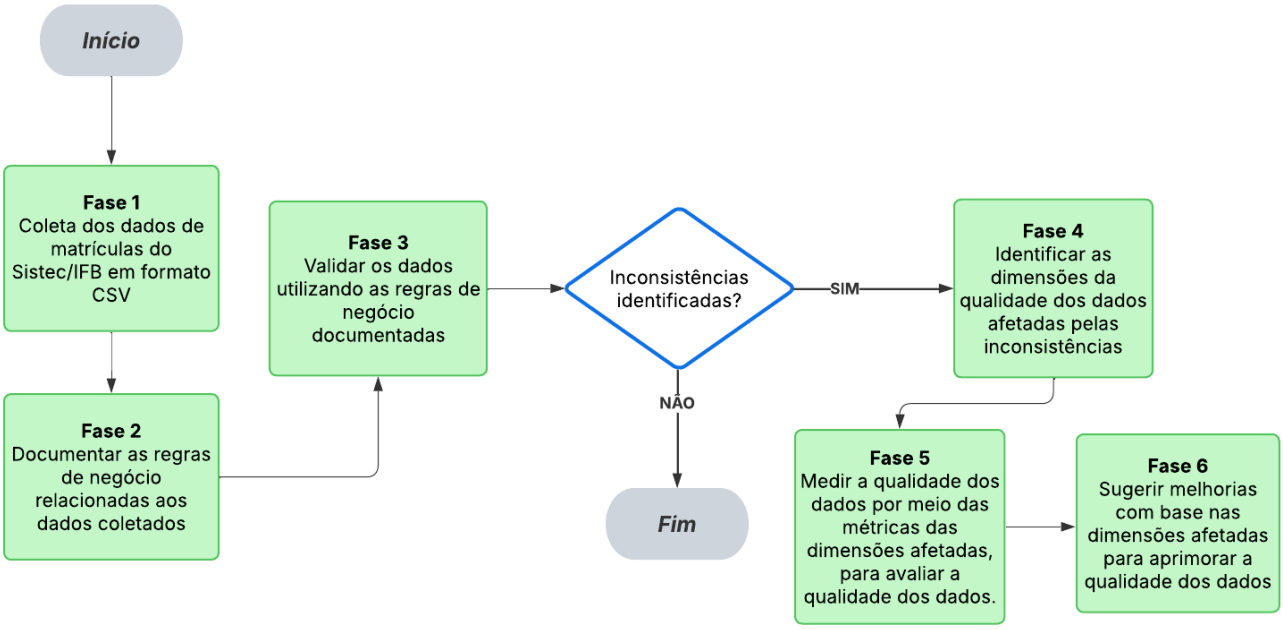
Após identificar as inconsistências e definir as dimensões afetadas com base nas irregularidades detectadas nos dados, são desenvolvidas funções específicas para calcular as métricas de cada dimensão, possibilitando uma avaliação detalhada da qualidade dos dados. Essa análise é essencial para verificar se os dados estão adequados à geração de informações relevantes e para planejar ações que promovam a melhoria de sua qualidade. Os resultados de cada métrica são apresentados em forma de porcentagem: quanto mais próximo de 100%, maior é a qualidade da dimensão analisada. Cada uma das fases mencionadas é representada no fluxograma que resume essas etapas. A Figura 4.4 apresenta esse fluxograma.

Quadro 4.3: Dimensões e métricas escolhidas para medir a qualidade dos dados

DQD	Descrição	Métricas
Exatidão	Mede a exatidão dos dados em relação à realidade ou o valor verdadeiro.	$\frac{\text{valores corretos}}{\text{total de valores}} \times 100$
Integridade	Precisão dos dados originais versus processados.	$\frac{\text{diferenças entre valores originais e processados}}{\text{total de valores}} \times 100$
Volatilidade	Tempo que os dados permanecem válidos.	$\frac{\text{criação - modificação}}{\text{atual - criação}} \times 100$
Acessibilidade	Facilidade de acesso e pesquisa dos dados.	$\frac{\text{valores acessíveis}}{\text{total de valores}} \times 100$
Segurança	Proteção dos dados por meio de políticas e criptografia.	-
Completude	Extensão dos dados completos.	$\frac{\text{valores não vazios}}{\text{total de valores}} \times 100$
Relevância	Importância dos dados para as análises.	$\frac{\text{acessos ao campo}}{\text{total de acessos}} \times 100$
Facilidade de Manipulação	Facilidade após pré-processamento.	$\frac{\text{diferenças na tabela}}{\text{total de valores}} \times 100$
Consistência	Coerência com tipos e esquemas definidos.	$\frac{\text{valores consistentes}}{\text{total de valores}} \times 100$
Legibilidade	Processamento sem erros semânticos.	$\frac{\text{valores sem erros}}{\text{total de valores}} \times 100$
Conformidade	Aderência a formatos definidos (e.g., datas, telefones).	$\frac{\text{valores com formato correto}}{\text{total de valores}} \times 100$
Unicidade	Redundância nos dados.	$\frac{\text{valores únicos}}{\text{total de valores}} \times 100$
Pontualidade	Atualização dos dados desde a última modificação.	$\frac{\text{atual - última modificação}}{\text{atual - criação}} \times 100$

Fonte: Adaptado de ELOUATAOUI et al, 2022.

Figura 4.4: Fluxograma das fases para a análise de qualidade dos dados



Fonte: Elaborado pela autora.

4.4 Fases para a melhoria na qualidade dos dados

Como citado no capítulo, Fases para analisar as inconsistências relatadas No tópico 3, Melhorias, após realizar as fases de avaliação e identificar as dimensões afetadas, ocorre a aplicação da fase de melhoria. Segundo os autores BATINI et al. (2009), a fase de melhoria é composta por etapas, estratégias e técnicas voltadas para a melhoria da qualidade dos dados.

Após identificar as dimensões afetadas pelas inconsistências nos dados do Sistec IFB e avaliar/medir a qualidade dos dados, foram seguidas etapas voltadas à melhoria da qualidade, conforme apresentado pelos autores BATINI et al. (2009) sobre a melhoria da qualidade dos dados. Essas etapas são:

- Avaliação de custos: análise dos custos, que podem ser diretos ou indiretos, relacionados à qualidade dos dados.
- Atribuição de responsabilidades de processo: definição dos responsáveis pelos processos que gerenciam os dados e do responsável direto pelos dados.
- Atribuição de responsabilidades de dados: identificação dos responsáveis e definição das atividades de produção e gestão dos dados.
- Identificação das causas dos erros: identifica as causas dos problemas de qualidades nos dados. Isso ajuda saber onde agir.
- Seleção de estratégias e técnicas: escolha de estratégias e técnicas adequadas para melhorar a qualidade dos dados, considerando os objetivos, o contexto e o orçamento disponível.
- Design de soluções de melhoria de dados: definição da solução com base na estratégia selecionada. A solução escolhida deve ser a mais eficiente e eficaz, composta por um conjunto de técnicas e ferramentas para aprimorar a qualidade dos dados.
- Controle de processo: definição de pontos de verificação nos processos de produção de dados, para garantir a manutenção da qualidade.
- Redesenho de processo: definição de ações de melhoria nos processos, visando aprimorar ainda mais a qualidade dos dados.
- Gestão de melhorias: estabelecimento de novas regras e diretrizes dentro da organização para assegurar a manutenção da qualidade dos dados.
- Monitoramento de melhorias: realização de atividades periódicas de monitoramento, com o objetivo de fornecer relatórios sobre os resultados do processo de melhoria e permitir ajustes dinâmicos.

Neste trabalho, as fases de melhoria não foram aplicadas diretamente aos dados do Sistec IFB, devido a limitações de acesso e de propriedade. Mesmo tratando-se de dados abertos, algumas etapas de melhoria exigem acessos mais específicos às bases de dados. As etapas e metodologias apresentadas servem como sugestões e guia para futuras melhorias, baseadas na validação realizada e nas dimensões afetadas dos dados de matrículas do Sistec IFB. Com base na metodologia *CDQ*, foram sugeridas melhorias.

Das etapas previstas, foram destacadas: a identificação das causas dos erros, por meio da análise das regras de negócio; a validação e detecção de irregularidades; o design de soluções para

melhoria; a escolha da metodologia *CDQ*, que abrange todas as fases desde a identificação de irregularidades até a avaliação e medição da qualidade e o controle do processo, com validações voltadas para erros de formato, preenchimento de campos obrigatórios, entre outros. Algumas etapas não foram sugeridas, pois exigem um maior nível de acesso ao armazenamento dos dados.

Para a melhoria da qualidade dos dados do Sistec IFB, foram sugeridas duas estratégias da metodologia *CDQ*: **baseadas em dados** e **baseadas em processos**. Segundo BATINI et al. (2009), as estratégias baseadas em dados atuam diretamente sobre o conteúdo dos dados, como, por exemplo, na atualização de valores desatualizados. Já as estratégias baseadas em processos visam à melhoria da qualidade por meio do redesenho de processos que criam ou modificam os dados, como no controle do formato antes do armazenamento.

Baseadas em dados

Segundo os autores BATINI et al. (2009), essas estratégias utilizam diversas técnicas, como algoritmos, heurísticas e atividades baseadas em conhecimento, para melhorar a qualidade dos dados. No caso dos dados do Sistec IFB, após as fases de reconstrução e avaliação/medição, foram seguidos alguns desses passos para a sugestão de estratégias orientadas por dados, que são:

1. Obtenção de novos dados: melhora a qualidade ao adquirir dados mais confiáveis, substituindo valores que apresentam problemas de qualidade.

2. Padronização (normalização): substitui ou completa valores que não estão em conformidade com as regras de formato estabelecidas.

3. Vinculação de registros: identifica registros em duas ou mais tabelas diferentes que se referem ao mesmo objeto do mundo real.

4. Integração de dados e esquemas: combina dados de várias fontes (como bancos de dados e arquivos), definindo uma visão unificada das informações fornecidas. Isso permite que os usuários acessem os dados de maneira integrada, independentemente da fonte original.

5. Confiabilidade da fonte: seleciona as fontes de dados e seus respectivos conteúdos com base na qualidade que apresentam.

6. Localização e correção de erros: identifica e elimina erros de qualidade nos dados, detectando registros que não atendem a um conjunto de regras previamente definidas.

7. Otimização de custo: define ações de melhoria da qualidade em um conjunto de dimensões, com o objetivo de reduzir erros e, consequentemente, os custos associados.

Essas técnicas de melhorias por estratégias orientadas a dados fornecem passos para aplicar melhorias nos dados, diretamente nos valores, substituindo erros e aumentando a qualidade.

Baseadas em processos

Como mencionado anteriormente, existem também as estratégias baseadas em processos. Os autores BATINI et al. (2009) destacam duas principais características estratégicas relacionadas aos processos. A primeira característica refere-se ao **controle de processo**, já mencionado entre as fases de melhoria listadas no início deste capítulo. O controle de processo inclui verificações e procedimentos de controle, que são:

1. Novos dados são criados.
2. Conjuntos de dados são atualizados.
3. Novos conjuntos de dados são acessados pelo processo, por exemplo, quando há integração com dados de outro sistema.

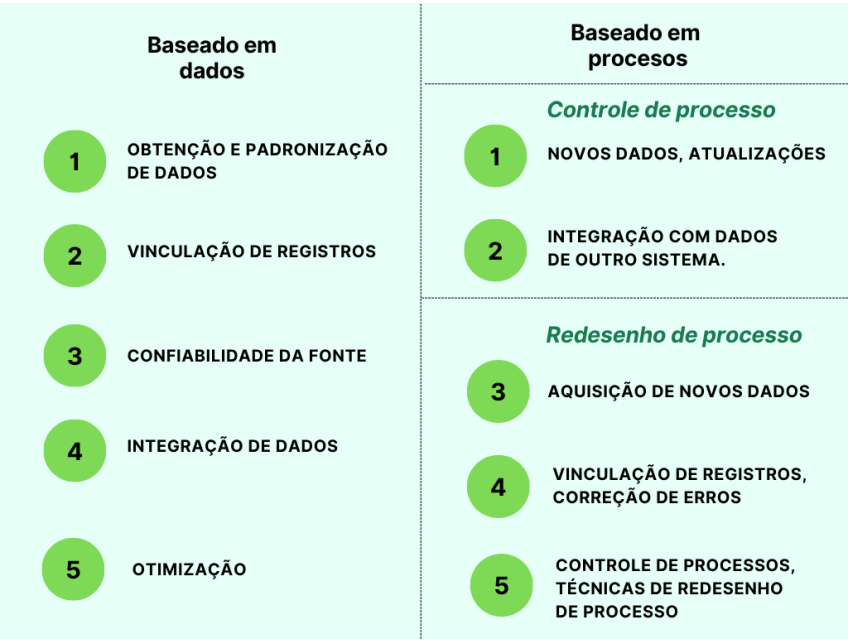
Essa estratégia é reativa e aplicada a eventos de modificação de dados; por isso, é considerada reativa, pois evita a desatualização dos dados e a perpetuação de erros. A segunda característica é o **redesenho de processo**, cujo objetivo é remover as causas da baixa qualidade dos dados e atualizar as atividades para que passem a produzir dados de melhor qualidade. Segundo os autores BATINI et al. (2009), existem diversas técnicas disponíveis, cada uma com seu custo. Essa fase de redesenho de processo é apresentada nas fases de melhoria listadas anteriormente, no início deste capítulo.

As ações descritas a seguir não serão aplicadas diretamente, mas são sugeridas como possíveis caminhos para melhorias, baseadas nas inconsistências identificadas nos dados do Sistec IFB. As sugestões seguem a metodologia de melhoria proposta pelos autores, servindo como referencial teórico e prático para possíveis intervenções futuras. O redesenho de processo modifica os processos com o objetivo de eliminar as causas da baixa qualidade dos dados e introduz novas atividades que geram dados de maior qualidade, que são:

1. Aquisição de novos dados.
2. Vinculação de registros (identificação de registros duplicados ou semelhantes).
3. Localização e correção de erros.
4. Controle de processos (evitar que dados incorretos sejam cadastrados).
5. Técnicas de redesenho de processos (modificação ou reestruturação de um processo).

A Figura 4.5 apresenta, de maneira resumida as técnicas explicadas acima baseadas em dados e em processos.

Figura 4.5: Técnicas para a melhoria da qualidade dos dados



Fonte: Elaborado pela autora.

Os autores BATINI et al. (2009) mencionam que, a longo prazo, as técnicas orientadas por processos apresentam desempenho superior às técnicas orientadas por dados, pois essa técnica vai na causa raiz dos problemas de baixa qualidade dos dados. No mais, essas técnicas podem ser extremamente caras no curto prazo. As estratégias orientadas por dados são relatadas como mais econômicas e eficientes a curto prazo, porém a longo prazo não sejam tão eficazes. Elas são adequadas para aplicações pontuais, mas não são recomendadas para dados que permanecem constantes durante a execução, ou seja, dados estáticos.

Para este trabalho, a metodologia *CDQ* abrange melhorias baseadas tanto em dados quanto em processos. Essa metodologia foi aplicada conforme a validação dos dados e a identificação das dimensões afetadas, apresentando, assim, sugestões para a melhoria da qualidade dos dados do Sistec IFB, após a identificação e análise das áreas que apresentam inconsistências e impactam a qualidade dos dados.

4.5 Métrica global para a qualidade dos dados

De acordo com MAKHOUL (2022), cada indicador de qualidade (DQ_u) possui uma métrica específica para medir a qualidade dos dados, conforme explicado na seção qualidade de dados. Portanto, é útil considerar uma métrica global que calcule a qualidade total. Para isso, atribui-se um peso a cada dimensão de qualidade, refletindo sua importância relativa.

O método sugerido para calcular a métrica global de qualidade dos dados em *Structural Health Monitoring (SHM)* envolve a combinação ponderada de diferentes indicadores de qualidade, como, por exemplo, exatidão, completude e consistência, entre outros, utilizados para analisar a qualidade dos dados.

O cálculo da métrica global de qualidade dos dados apresentado por MAKHOUL (2022) (DQ_{Total}) é realizado pela soma ponderada dos valores dos indicadores de qualidade. é realizado por meio da soma ponderada dos valores dos indicadores de qualidade. Cada indicador recebe um peso (ω_u) que reflete sua importância, e a soma de todos os pesos deve ser igual a 1 ($\sum \omega_u = 1$).

A fórmula para calcular o DQ_{Total} é:

$$DQ_{Total} = \sum(\omega_u \cdot DQ_u)$$

Onde:

- ω_u é o peso atribuído ao indicador.
- DQ_u é a métrica de qualidade do indicador.

Este método permite uma avaliação global da qualidade dos dados, considerando a importância relativa de cada indicador na composição da qualidade total.

MAKHOUL (2022) também apresenta uma tabela que contém os rótulos de qualidade dos dados e as faixas de limiar correspondentes para a métrica global DQ_{Total}. Cada faixa de limiar classifica a qualidade dos dados em diferentes níveis. A seguir, essa tabela é apresentada em formato adaptado no quadro 4.4.

A qualidade dos dados é classificada de acordo com os seguintes rótulos e faixas de limiar para a métrica global. DQ_{Total}:

- Excelente: quando a métrica de qualidade dos dados está entre 0,8 e 1, a qualidade dos dados é considerada excelente, refletindo dados altamente confiáveis e completos.
- Bom: para valores de DQ entre 0,6 e 0,8, a qualidade dos dados é boa.

Quadro 4.4: Rótulos de qualidade de dados e intervalos de limiar para a métrica global

Rótulo de Qualidade	Faixa de Limiar de DQ
Excelente	$0{,}8 \leq \text{DQ} \leq 1$
Bom	$0{,}6 \leq \text{DQ} < 0{,}8$
Médio	$0{,}4 \leq \text{DQ} < 0{,}6$
Fraco	$0{,}2 \leq \text{DQ} < 0{,}4$
Muito Fraco	$0 \leq \text{DQ} < 0{,}2$

- Médio: se a métrica *DQ* estiver entre 0,4 e 0,6, os dados são considerados de qualidade média.
- Fraco: com *DQ* entre 0,2 e 0,4, os dados apresentam qualidade fraca.
- Muito fraco: quando *DQ* está entre 0 e 0,2, os dados são classificados como de qualidade muito fraca.

Esses rótulos e limiares fornecem uma maneira clara de categorizar a qualidade dos dados em um *SHM*, auxiliando na identificação das áreas que necessitam de melhorias. Para apresentar os resultados da métrica global da análise da qualidade dos dados do Sistec IFB, os rótulos e limiares foram considerados.

4.6 Aplicação das fases para analisar a qualidade dos dados

Nesta seção, são aplicadas as fases de reconstrução e avaliação/medição apresentadas no capítulo de metodologia. Cada fase listada foi seguida para realizar a análise da qualidade dos dados do Sistec IFB.

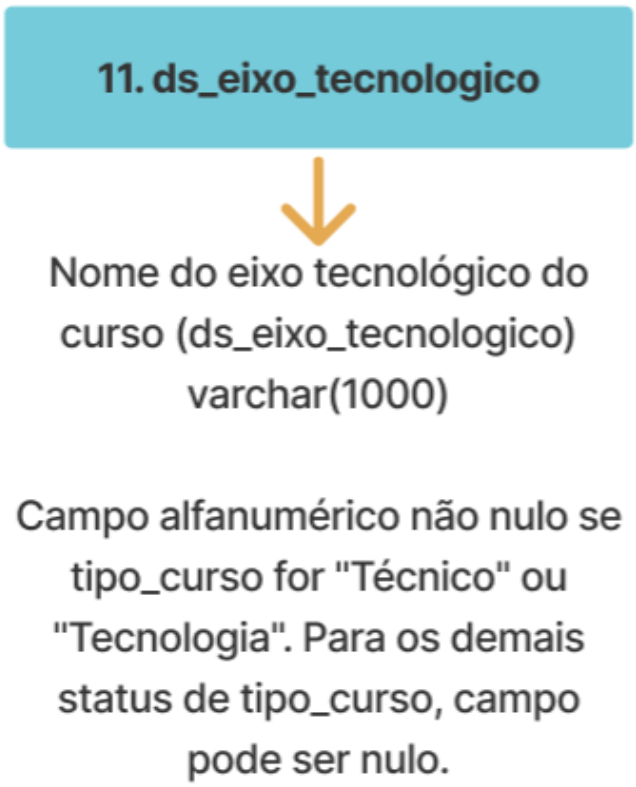
4.6.1 Reconstrução

Na primeira etapa, referente à reconstrução, foram elaboradas documentações detalhadas sobre os dados do Sistec ao longo do projeto de pesquisa mencionado, descrevendo cada dado e suas respectivas regras de governança.

Essa fase de reconstrução é fundamental, pois a existência dessas documentações facilita o entendimento das informações, a validação dos dados coletados e a verificação da conformidade com as regras estabelecidas. Ter registros que descrevam os dados, seus tipos e o significado de cada informação contribui diretamente para o desenvolvimento e a execução adequada das fases seguintes do projeto.

Essas documentações foram produzidas com base nos princípios de governança de dados, contendo, de forma detalhada, as regras que os dados devem seguir no momento da coleta e do armazenamento. Entre essas regras estão: o tipo do dado; regras de consistência como, por exemplo, a data de início da matrícula não poder ser posterior à data de fim da matrícula; a possibilidade ou não de o dado estar em branco; e informações descritivas sobre os dados, como o nome da coluna e o conteúdo que ela armazena. A Figura 4.6 apresenta essa documentação de um dos registros pertencentes à coluna dos dados. Todas as colunas foram descritas conforme o modelo apresentado na figura.

Figura 4.6: Documentação da regra de negócios de uma das colunas



Fonte: Elaborado pela autora.

Essas documentações são fundamentais para compreender as características de cada dado, utilizando-as para verificar possíveis irregularidades. Neste trabalho, foram utilizados dados do Sistec referentes ao IFB. O quadro 4.5 apresenta informações extraídas do dicionário de dados disponível no portal de dados abertos do Sistec IFB, relacionadas às matrículas do IFB.

Para realizar a análise da qualidade dos dados, foram utilizados o dicionário de dados apresentado no quadro 4.6 e as documentações descritas na Figura 4.6. Essas documentações fornecem detalhes sobre os dados, permitindo uma compreensão mais completa para a validação e identificação das dimensões de qualidade afetadas por inconsistências. Com isso, torna-se possível medir a qualidade dos dados de forma mais precisa, direcionando ações de melhoria conforme as dimensões afetadas por irregularidades.

As colunas ano, período e cadastro_matricula não foram verificadas por não serem consideradas relevantes para a análise da qualidade dos dados. As demais colunas, entretanto, foram devidamente analisadas.

4.6.2 Avaliação/Medição

Nesta seção, apresenta-se o desenvolvimento das fases de avaliação e medição mencionadas na seção fases para analisar as inconsistências relatadas, que tratam da análise das inconsistências identificadas.

- Análise de dados: conforme explicado na seção fases para analisar as inconsistências relatadas, foi realizada uma documentação detalhada dos dados, com foco em suas regras de negócio, como dados que podem ser nulos, regras de consistência do banco de dados, entre outros. A partir disso, foram criadas documentações que orientam a validação dos dados. Para este trabalho, essas documentações foram aplicadas aos dados abertos do Sistec IFB.

- Análise dos requisitos de qualidade dos dados: com as regras de negócio definidas, foram desenvolvidos códigos para verificar a conformidade dos dados. Um exemplo é a validação dos registros da coluna "nome do eixo tecnológico", conforme a regra descrita no capítulo aplicação das fases para analisar a qualidade dos dados na parte de reconstrução. Quando inconsistências são identificadas, o código gera um arquivo CSV contendo os registros que não estão conforme a regra de negócio dos dados. Todas as colunas presentes no dicionário de dados tiveram suas respectivas regras documentadas e validadas por meio de um *script* desenvolvido na linguagem Python. O algoritmo 1, apresentado a seguir, exemplifica esse processo ao verificar os valores da coluna nome do eixo tecnológica.

Algorithm 1 Validação da coluna `ds_eixo_tecnologico`

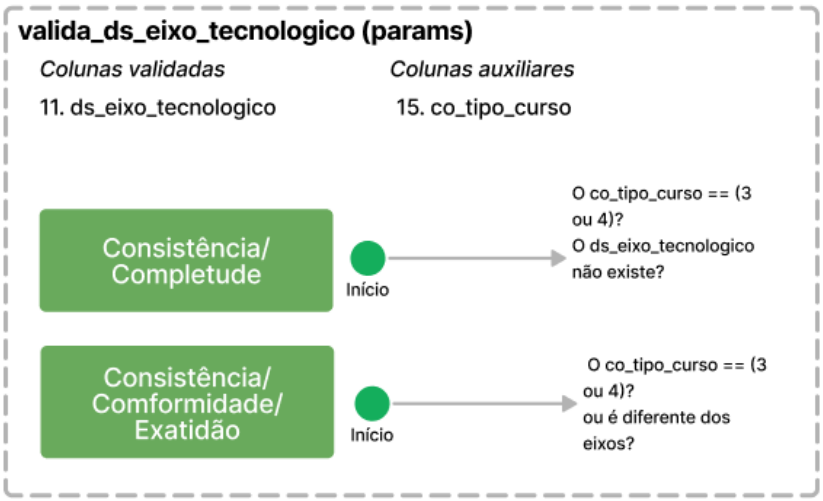
```
1: Procedimento VALIDAEIXOTECNOLOGICO(df_colunas, base_info)
2:     inicializar DataFrame vazio resultado
3:     definir lista de eixos tecnológicos válidos
4:     definir tipos de curso que exigem eixo tecnológico: 3 e 4
5:     obter categoria e código da coluna ds_eixo_tecnologico
6:     para todo linha em df_colunas faça
7:         extrair codigo_universal_linha, co_tipo_curso e conteudo
8:         se co_tipo_curso está em {3, 4} então
9:             se conteudo está vazio ou nulo então
10:                adicionar inconsistência "vazio_não_conforme_a_regra" em resultado
11:            senão se conteudo não está na lista de eixos tecnológicos válidos então
12:                adicionar inconsistência "diferente_dos_eixos_e_não_conforme_a_regra" em resultado
13:            fim se
14:        fim se
15:    fim para
16:    se resultado não está vazio então
17:        gerar arquivo CSV com inconsistências
18:        limpar resultado
19:    senão
20:        exibir mensagem: "Nenhuma inconsistência encontrada"
21:    fim se
22: fim Procedimento
```

- Identificação de áreas críticas: todos os dados de todas as colunas presentes no dicionário de dados foram analisados, validados e tiveram sua qualidade avaliada.

- Medição da qualidade: com base nas regras de negócio dos dados, foram definidas as dimensões objetivas a serem avaliadas. Já as dimensões subjetivas não puderam ser avaliadas, pois requerem um conhecimento mais aprofundado sobre o armazenamento e o controle realizados pelos administradores dos dados. As dimensões objetivas foram avaliadas conforme as regras dos dados, como, por exemplo, dado não pode estar vazio, tipo de dado inadequado, conforme ilustrado na Figura 4.6, que apresenta uma das documentações das colunas dos dados.

A Figura 4.7 apresenta uma das documentações dos dados de uma das colunas do Sistec IFB, mostrando como a categorização foi desenvolvida para a documentação da coluna ds_eixo_tecnológico, conforme ilustrado na Figura 4.6. Para cada dado da coluna do Sistec IFB, cada valor segue suas próprias regras de negócio e possui uma categorização específica das dimensões de qualidade afetadas. Essas dimensões foram identificadas e categorizadas de acordo com os dados analisados. Após essa categorização, foram desenvolvidos *scripts* na linguagem Python que calculam métricas específicas para cada dimensão, com o objetivo de medir a qualidade dos dados.

Figura 4.7: Categorização das dimensões de acordo com o dicionário e as regras de negócio de uma das colunas



Fonte: Elaborado pela autora.

O algoritmo 2, apresentado a seguir, exemplifica a métrica de completude, que mede uma das dimensões mostradas na Figura 4.7. Cada dimensão de qualidade possui uma métrica específica para avaliar a qualidade dos dados. Para as dimensões objetivas, foram desenvolvidos *scripts* na linguagem Python que implementam essas métricas. As dimensões subjetivas, explicadas anteriormente, não foram analisadas, pois requerem um conhecimento mais aprofundado sobre a administração dos dados. As dimensões objetivas afetadas tiveram seus respectivos códigos criados e aplicados para medir a qualidade dos dados nos registros onde foram identificadas inconsistências.

Algorithm 2 Cálculo da métrica de completude

Procedimento METRICA_COMPLETUDE(df_dados_categorizados, total_df_sistec)
    completude ← filtrar `df_dados_categorizados` onde `dimensão = "Completude"`
    completude_total_linhas ← número de linhas de `completude`
    imprimir “Total de inconsistência de completude:”, completude_total_linhas
    total_linhas_diferenca ← total_df_sistec − completude_total_linhas
    imprimir “Diferença de total inconsistência / Total linhas dataframe:”, total_linhas_diferenca
    porcentagem_qualidade ← (total_linhas_diferenca / total_df_sistec) × 100
    porcenta_duas_decimais ← formatar porcentagem_qualidade com 3 casas decimais
    imprimir “porcenta_duas_decimais”
    salvar_dados_metrica(“Completude”, porcenta_tres_decimais)
fim Procedimento

[1]

[1]O código-fonte desenvolvido está disponível no GitHub. Disponível em: `https://github.com/infocbra/infocbra-tcc-qualidade-dados-sistec-ifb`

Quadro 4.5: Dicionário dos dados do Sistec IFB

Dado	Nome do Campo	Tipo	Descrição
Aluno	aluno	Alfanumérico	Identificação anônima do aluno
Código do Ciclo de Matrícula	co_ciclo_matricula	Numérico	Código de identificação do ciclo de matrícula ao qual o aluno pertence
Unidade de Ensino (Campus)	unidade_ensino	Texto	Unidade de ensino (campus) ao qual o aluno está vinculado
Dependência Administrativa	dependencia_adm	Texto	Tipo de dependência administrativa ao qual o aluno está vinculado (Pública ou Privada)
Sistema de Ensino	sistema_ensino	Texto	Sistema de ensino ao qual o aluno está vinculado (municipal, estadual, federal)
Município	municipio	Texto	Município no qual o aluno está matriculado
Estado	estado	Texto	Estado no qual o aluno está matriculado
Período	periodo	Alfanumérico	Período de duração do ciclo de matrícula
Data de Início	dt_data_inicio	Alfanumérico	Data do início do ciclo de matrícula
Data do Fim Previsto	dt_data_fim_previsto	Alfanumérico	Data do fim previsto do ciclo de matrícula
Nome do Eixo Tecnológico	ds_eixo_tecnologico	Texto	Nome do eixo tecnológico ao qual o curso pertence
Código do Curso	co_curso	Numérico	Código identificador do curso
Nome do Curso	no_curso	Texto	Nome do curso
Data de Deferimento do Curso	dt_deferimento_curso	Alfanumérico	Data de deferimento do curso
Código do Tipo do Curso	co_tipo_curso	Numérico	Código identificador do tipo do curso
Tipo do Curso	tipo_curso	Texto	Nome do tipo do curso
Código do Tipo do Nível do Curso	co_tipo_nivel	Numérico	Código identificador do tipo do nível do curso
Tipo do Nível do Curso	ds_tipo_nivel	Texto	Nome do tipo do nível do curso
Nome do Ciclo	nome_ciclo	Texto	Nome do ciclo de matrícula
Data de Cadastro do Ciclo	dt_cadastro_ciclo	Alfanumérico	Data do cadastro do ciclo de matrícula
Carga Horária	carga_horaria	Numérico	Carga horária do ciclo de matrícula
Data de Cadastro do Aluno no Sistema	dt_cadastro_aluno_sistema	Alfanumérico	Data de cadastro do aluno no sistema
Período de Cadastro da Matrícula do Aluno	periodo_cadastro_matricula_ano	Alfanumérico	Período de cadastro do aluno no sistema (mês e ano)
Modalidade de Pagamento	modalidade_pagto	Texto	Modalidade de pagamento do curso
Situação da Matrícula	situacao_matricula	Texto	Situação da matrícula
Tipo de Cota	tipo_cota	Texto	Tipo de cota da matrícula
Atestado de Baixa Renda	atestado_baixarenda	Texto	Identificador de atestado de baixa renda da matrícula
Tipo de Oferta	tipo_oferta	Texto	Tipo de oferta do curso
Ano	ano	Numérico	Ano da matrícula
Modalidade de Ensino	modalidade_ensino	Texto	Modalidade de ensino do curso
Data de Ocorrência do Ciclo	dt_ocorrencia_ciclo	Alfanumérico	Data de ocorrência do ciclo de matrícula
Data de Ocorrência da Matrícula	dt_ocorrencia_matricula	Alfanumérico	Data de ocorrência da matrícula
Data da Última Alteração	data_ultima_alteracao	Alfanumérico	Data de ocorrência da última alteração da situação da matrícula
Vagas Ofertadas	vagas_ofertadas	Numérico	Quantidade de vagas ofertadas no ciclo
Total de Inscritos	total_inscritos	Numérico	Quantidade de inscritos no ciclo
Código do Status da Matrícula	co_status_matricula	Numérico	Código do status ou situação da matrícula
Sexo	sg_sexo	Texto	Sigla do sexo do aluno
Data de Nascimento	dt_data_nascimento	Alfanumérico	Data de nascimento do aluno
Nome Completo do Agrupador	nome_completo_agrupador	Texto	Nome completo do Instituto Federal ao qual a matrícula está vinculada
Sigla do Agrupador	sigla_agrupador	Texto	Sigla do Instituto Federal ao qual a matrícula está vinculada

Fonte: https://diretorios.ifb.edu.br/diretorios/1558

5
Análise da qualidade dos dados

Neste capítulo é apresentada a análise da qualidade dos dados, após a aplicação de todas as fases da metodologia.

5.1 Inconsistências identificadas nos dados

Para medir a qualidade dos dados, foram aplicadas as etapas das fases descritas no capitulo Metodologia na seção fases para analisar as inconsistências relatadas. A metodologia utilizada foi a *CDQ*.

O quadro 5.1 apresenta as inconsistências encontradas nos dados do Sistec IFB após a validação, utilizando as regras de negócio desenvolvidas na fase de reconstrução. Todas as colunas do dicionário apresentadas no quadro 4.5 tiveram suas regras de negócio verificadas por meio de *script*, conforme explicado na seção avaliação/medição.

Quadro 5.1: Inconsistências relatadas nos dados do Sistec IFB

Coluna	Tipo de Inconsistência	Total
dt_cadastro_ciclo	dt_cadastro_ciclo_inferior_não_conforme_a_regra	46.352
dt_cadastro_aluno_sistema	dt_cadastro_aluno_sistema_inferior_não_conforme_a_regra	18.044
tipo_oferta	diferente_hifen_e_não_conforme_a_regra	15.856
no_curso	caracter_especial	120
nome_ciclo	caracter_especial	120
-	duplicidade	24
sg_sexo	valor_diferente_F_M	3
	vazio	1
tipo_cota	vazio	1
Soma total de inconsistências		**80.521**

Fonte: Elaborado pela autora a partir da análise de qualidade dos dados do Sistec IFB.

Após a validação dos dados com suas respectivas regras de negócio, as inconsistências foram categorizadas de acordo com as dimensões da qualidade dos dados. Irregularidades nos dados afetam diretamente essas dimensões, conforme apresentado anteriormente na seção avaliação/medição. Cada verificação que não está de acordo com a regra tem sua dimensão de qualidade correspondente identificada. O quadro 5.2 apresenta essa categorização com base na dimensão afetada pela irregularidade.

Compreender qual dimensão é afetada pelas irregularidades nos dados é fundamental para medir a qualidade. Na metodologia, são apresentadas as definições de cada dimensão e a forma de calculá-las, conforme mostrado no quadro 4.3, sendo que cada uma possui sua própria métrica de avaliação.

Devido às limitações deste trabalho, foram analisadas cinco dimensões: unicidade, exatidão, completude, conformidade e consistência. Essas dimensões foram escolhidas por

Quadro 5.2: Categorização das inconsistências por dimensão de qualidade

Coluna	Tipo de Inconsistência	Dimensão	Total
dt_cadastro_ciclo	dt_cadastro_ciclo_inferior_não_conforme_a_regra	Consistência	46.352
dt_cadastro_aluno_sistema	dt_cadastro_aluno_sistema_inferior_não_conforme_a_regra	Consistência	18.044
tipo_oferta	diferente_hifen_e_não_conforme_a_regra	Consistência	15.856
no_curso	caracter_especial	Conformidade/Exatidão	120
nome_ciclo	caracter_especial	Conformidade/Exatidão	120
-	duplicidade	Unicidade	24
sg_sexo	valor_diferente_F_M	Conformidade/Exatidão	3
	vazio	Completude	1
tipo_cota	vazio	Completude	1
Soma total de inconsistências			**80.521**

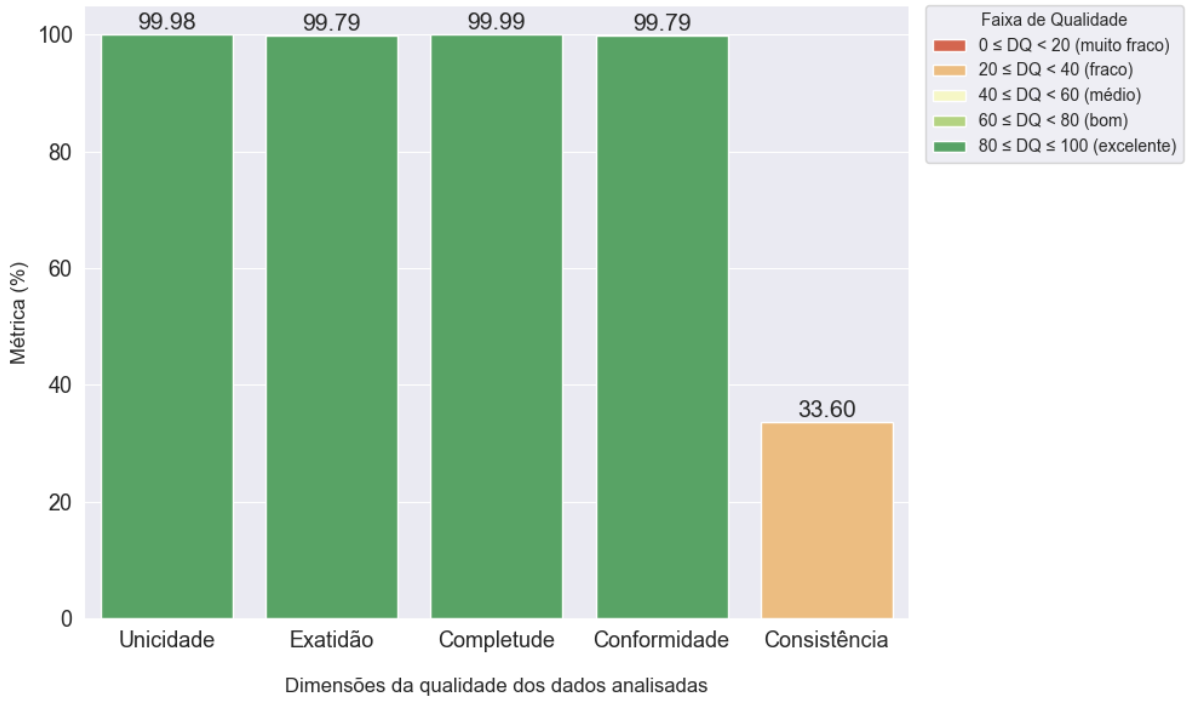
Fonte: Elaborado pela autora a partir da análise de qualidade dos dados do Sistec IFB.

serem objetivas e por não exigirem conhecimento administrativo mais aprofundado sobre os dados, como frequência de atualização, estrutura do banco de dados, entre outros fatores mais complexos de se analisar neste contexto. Apesar de os dados utilizados serem abertos, há limitações quanto ao conhecimento sobre seu gerenciamento. Por isso, este trabalho focou nessas quatro dimensões, que, como mostrado no capítulo de revisão da literatura no artigo *Big data quality: a quality dimensions evaluation* estão entre as mais utilizadas na avaliação da qualidade dos dados.

5.2 Resultado das métricas de dimensão da qualidade dos dados

A Figura 5.1 a seguir apresenta um gráfico com o resultado das métricas que medem a qualidade dos dados nas cinco dimensões avaliadas. Quanto mais próximo de 100%, melhor é a qualidade da dimensão analisada.

Figura 5.1: Gráfico do resultado das métricas de dimensão de qualidade de dados



Fonte: Elaborado pela autora a partir da análise de qualidade dos dados do Sistec IFB.

O gráfico acima apresenta os resultados em porcentagem. O eixo Y representa os valores de 0 a 100%, enquanto o eixo X corresponde às dimensões da qualidade dos dados avaliadas. As quatro dimensões analisadas são explicadas na seção de metodologia, juntamente com as métricas utilizadas para medir cada uma. A dimensão de unicidade verifica a presença de registros duplicados, avaliando o quanto os dados são únicos. A dimensão de exatidão verifica se os valores armazenados estão corretos e de acordo com os valores esperados, previamente conhecidos. A dimensão de completude analisa se os campos obrigatórios estão devidamente preenchidos, indicando a ausência de dados faltantes. A dimensão de conformidade verifica se os dados estão no padrão esperado, como data, valores que eram para serem numéricos ou texto. A dimensão de consistência é responsável por avaliar se os dados seguem as regras de negócio estabelecidas e se mantêm coerência entre diferentes campos, conforme esperado no contexto do banco de dados.

Com o resultado de cada métrica das dimensões analisadas, é possível compreender a qualidade dos dados. Essas quatro métricas são fundamentais para garantir essa qualidade. A partir da análise realizada nos dados do Sistec IFB, torna-se possível identificar o nível de qualidade dos dados em relação a essas dimensões, oferecendo uma visão mais clara e possibilitando identificar áreas que podem precisar de melhorias para aumentar a qualidade dos dados.

Dessa forma, os resultados obtidos pela aplicação das métricas permitem uma avaliação objetiva da qualidade dos dados do Sistec IFB, destacando o que precisa ser aprimorado ou mantido. A análise baseada nas dimensões de unicidade, exatidão, completude, conformidade e consistência fornece informações essenciais para a tomada de decisões voltadas à melhoria dos processos de gestão e utilização dos dados, além de avaliar o nível de qualidade que os dados possuem.

5.3 Resultado da métrica global

Como apresentado no capítulo métrica global, MAKHOUL (2022) apresenta em seu trabalho uma fórmula que calcula a qualidade total dos dados, atribuindo um peso a cada dimensão de qualidade utilizada na avaliação.

Conforme apresentado nos resultados acima, foram consideradas cinto métricas: unicidade, exatidão, completude conformidade e consistência. A métrica global atribui um peso para cada dimensão, de acordo com sua importância relativa para a qualidade dos dados. A soma dos pesos deve ser igual a 1, representando 100%, que é o valor máximo que cada métrica de dimensão pode alcançar.

A fórmula apresentada para o cálculo do DQ_{Total} é:

$$DQ_{\text{Total}} = \sum(\omega_u \cdot DQ_u)$$

Como foram utilizadas cinco métricas para avaliar a qualidade dos dados, o valor 1 é dividido entre as cinco dimensões, recomenda-se que o peso atribuído a cada dimensão seja proporcional à sua importância no contexto dos dados. Para os dados do Sistec IFB, onde foram aplicadas cinco métricas, os pesos foram ajustados conforme a relevância de cada dimensão, garantindo que a soma total seja igual a 1. Dessa forma, cada dimensão recebeu um peso diferente, refletindo sua importância relativa no conjunto de dados analisados.

- Unicidade (0,1): importante para evitar valores repetidos do mesmo aluno, o que prejudica identificar qual informação é a mais atual.

- Exatidão (0,3): importante para garantir que os dados acadêmicos estejam corretos, pois a exatidão das informações, como as de matrículas, é fundamental.
- Completude (0,25): essencial para assegurar que todos os campos obrigatórios estejam preenchidos, evitando a falta de informações que são obrigatórias o preenchimento.
- Conformidade (0,15): necessária para garantir que os dados estejam em conformidade com normas e formatos exigidos, como datas, campos numéricos ou textos.
- Consistência (0,2): importante para manter a integridade dos dados entre as tabelas, regras do banco e a coerência dos dados.

Aqui, DQ_u representa o resultado da métrica para cada dimensão. Na Figura 5.1, onde é apresentado um gráfico, cada dimensão apresenta seu resultado em porcentagem: unicidade 99,98%, exatidão 99,79%, completude 99,99%, conformidade 99,79%, consistência 33,60%. Aplicando a fórmula aos valores:

$$\begin{aligned} DQ_{\text{Total}} &= (0{,}1 \times 99{,}98) + (0{,}3 \times 99{,}79) + (0{,}25 \times 99{,}99) \\ &\quad + (0{,}15 \times 99{,}79) + (0{,}2 \times 33{,}60) \\ &= 9{,}998 + 29{,}937 + 24{,}9975 + 14{,}9685 + 6{,}72 \\ &= 86{,}621 \\ &\approx \mathbf{86{,}62}\% \end{aligned}$$

Portanto, o resultado da métrica global para a qualidade total dos dados do Sistec IFB é de aproximadamente 86,62%. No capítulo métrica global, com base em MAKHOUL (2022), é apresentada uma tabela que classifica os resultados da fórmula, considerando-os excelentes quando $0.8 \leq \text{DQ} \leq 1$. Dessa forma, a análise indica que o Sistec IFB possui uma qualidade excelente, segundo as faixas de classificação da métrica global.

6
Sugestões de melhorias para aumentar a qualidade dos dados

Este capítulo apresenta sugestões de melhorias baseadas em dados e em processos. Como mencionado na seção fases para a melhoria, BATINI et al. (2009) menciona duas abordagens principais: a **baseada em dados** e a **baseada em processos**, as duas estratégias contém etapas que foram utilizadas para propor sugestões voltadas à melhoria da qualidade dos dados do Sistec IFB, com base nos resultados obtidos após a analise da qualidade dos dados das cinco dimensões que são: unicidade, exatidão, completude, conformidade e consistência.

A aplicação dessas estratégias apresentadas pelos autores BATINI et al. (2009) permite atuar de forma orientada na melhoria da qualidade dos dados. As duas abordagens têm efeitos diferentes: uma é direcionada aos dados e outra ao controle e redesenho, que inclui verificações e alterações no processo de cadastro e gestão dos dados. No quadro 6.1 é utilizada como orientação para sugerir as duas abordagens, utilizando as inconsistências relatadas nos dados e as dimensões afetadas para propor ações de melhoria nos dados do Sistec IFB.

Quadro 6.1: Categorização das inconsistências por dimensão de qualidade

Coluna	Tipo de Inconsistência	Dimensão	Total
dt_cadastro_ciclo	dt_cadastro_ciclo_inferior_não_conforme_a_regra	Consistência	46.352
dt_cadastro_aluno_sistema	dt_cadastro_aluno_sistema_inferior_não_conforme_a_regra	Consistência	18.044
tipo_oferta	diferente_hifen_e_não_conforme_a_regra	Consistência	15.856
no_curso	caracter_especial	Conformidade/Exatidão	120
nome_ciclo	caracter_especial	Conformidade/Exatidão	120
-	duplicidade	Unicidade	24
sg_sexo	valor_diferente_F_M	Conformidade/Exatidão	3
	vazio	Completude	1
tipo_cota	vazio	Completude	1
Soma total de inconsistências			**80.521**

Fonte: Elaborado pela autora a partir da análise de qualidade dos dados do Sistec IFB.

6.1 Abordagem baseada em dados

A abordagem **baseada em dados** refere-se à atuação direta no conteúdo já armazenado, por meio de ações que incluem: obtenção de novos dados, padronização (normalização), vinculação de registros, integração de dados e esquemas, confiabilidade, localização e correção de erros, e otimização de custos. A seguir, cada uma dessas ações será sugerida para os dados do Sistec IFB.

6.1.1 Obtenção de novos dados

Na dimensão **completude**, acontece a irregularidades na falta de preenchimento de campos obrigatórios. Após a validação dos dados do Sistec IFB, foram encontrados dados faltantes, como apresentado no quadro 6.1, onde o tipo da inconsistência aparece como vazio.

Apesar de grande parte dos dados estarem preenchidos, é necessário aplicar melhorias. As sugestões são:

- Buscar registros anteriores do aluno no Sistec vinculados a outras instituições para recuperar dados ausentes.
- Utilizar o e-mail para contato direto e coleta das informações faltantes.
- Consultar cadastro de matriculas do aluno em ciclos de matriculas anteriores para reaproveitar dados já registrados preenchendo dados ausentes.

6.1.2 Padronização (ou normalização)

Para as dimensões de **exatidão** e **conformidade**:

- Padronizar os valores com base em referências conhecidas. Por exemplo, substituir caracteres especiais no início do nome da instituição pelo nome correto, e corrigir valores inconsistentes no campo sexo do aluno, padronizar o formato das datas.

6.1.3 Vinculação de registros

A vinculação de registros é útil para a dimensão de **unicidade**, pois permite identificar registros com as mesmas informações e duplicações, mantendo apenas a última atualização. Isso evita repetições desnecessárias. No quadro 6.1 que é a analise da qualidade dos dados do Sistec IFB contém valores repetidos.

6.1.4 Integração de dados e esquemas

Uma forma de melhorar a qualidade relacionada à **completude** é aproveitar a estrutura já existente no Sistec, que integra dados de matrículas de todo o Brasil. Essa integração facilita a consulta por parte de usuários e instituições. Dessa forma é possível ter ações:

- Integrar dados de outros sistemas ou instituições para preencher dados faltantes que eram obrigatórios, contribuindo para as dimensões de **completude**.

6.1.5 Confiabilidade da fonte

Dados duplicados ou com problemas de consistência afetam a confiabilidade. Aprimorar as dimensões de **unicidade** e **consistência** entre registros aumenta a confiança nas informações. A consistência verifica se os dados estão duplicados com o mesmo conteúdo ou se violam regras do banco de dados. Como mostrado no quadro 6.1, existem registros em que a data de cadastro do aluno no sistema é inferior à data de cadastro do ciclo, o que não deveria ocorrer, afetando assim a consistência dos dados. Esse tipo de inconsistência é resolvido de forma mais eficiente com base em processos. Porém, uma sugestão para a confiabilidade é utilizar dados que seguem as regras de negócio para extrair informações e gerar visualizações, após validar e identificar os dados que estão seguindo as regras de negócio.

6.1.6 Localização e correção de erros

Como realizado no capitulo aplicação das fases para analisar a qualidade dos dados na seção avaliação/medição e apresentado no quadro 6.1, foram identificadas as dimensões afetadas, e o gráfico 5.1 mostra os resultados das métricas que medem a porcentagem da qualidade para cada dimensão. A partir disso, é possível:

- Sugerir melhorias direcionadas às dimensões afetadas, possibilitando entender quando e onde a irregularidade ocorreu, para aplicar ações corretivas nos dados. Com a análise dos dados do Sistec IFB, é possível propor sugestões de melhoria mais precisas, pois os erros foram identificados.

- Corrigir data inconsistentes, valores muito diferentes do esperado, então localizar o erro e corrigir.

6.1.7 Otimização de custo

Utilizar as análises dos dados e a identificação das dimensões afetadas contribui para a otimização de custos, como por exemplo:

- Escolher quais erros vale mais a pena corrigir com base no custo/benefício.

- Aplicar melhorias diretamente nas dimensões afetadas, otimizando tempo e recursos. Saber onde está o erro facilita onde direcionar ações corretivas. Por exemplo, nas dimensões de **exatidão** e **conformidade**, é possível ajustar dados que não seguem o padrão das regras de negócio, localizar valores ausentes e coletar essas informações. Entender onde estão os problemas otimiza o tempo para resolver.

6.2 Abordagem baseada em processos

A abordagem de dados é uma estrategia que contém duas características citadas pelos autores BATINI et al. (2009) que são: **controle de processo** e **redesenho de processo**, ambas são direcionadas para melhoria da coleta dos dados, restrutura do desenho que coleta os dados, regras que fazem parte da coleta e gerenciamento, essa a longo prazo é a mais indicada, pois resolve o problema, sem a persistência ao coletar os dados e ter que alterar os dados para ajustar irregularidades. Controle de processo insere verificações e procedimentos de controle no processo de produção de dados quando:

6.2.1 Controle de processo:

- Novos dados são criados

- Conjuntos de dados são atualizados.

- Novos conjuntos de dados são acessados, pelo processo quando existe integração de dados.

Sugestões de controle de processo contém etapas que são sugeridas para os dados do Sistec IFB que serve para: unicidade, completude, consistência, exatidão, conformidade, essas etapas são:

- Para a dimensão de **unicidade**, utilizar o preenchimento das colunas co_matricula, junto com co_aluno e co_ciclo_matricula, para verificar se o aluno já está matriculado naquela oferta de curso, evitando dados duplicados. As colunas co_matricula e co_aluno também existem, mas não foi possível ter acesso a elas por se tratarem de dados sensíveis. No momento do cadastro, deve-se verificar se essas três colunas já possuem registros cadastrados, para evitar duplicidade de dados.

- Para a dimensão de **conformidade**, deve-se verificar se o tipo do dado, no momento do cadastro, está conforme as regras estabelecidas para aquele campo, evitando inconsistências em padrões de data, e-mail ou outros dados que exigem formatos e tipos específicos, sejam eles numéricos ou textuais.

- Para **completude** não permitir campos obrigatórios sem preenchimento.

6.2.2 Redesenho de processo

O objetivo é alterar o processo de trabalho para garantir que os dados corretos sejam produzidos conforme as regras desde o início do cadastro. A seguir, são sugeridas etapas específicas para os dados do Sistec IFB, baseadas nas dimensões analisadas:

- Para a dimensão de **exatidão**, os dados que devem ser preenchidos com valores previamente conhecidos, como o nome da instituição, sigla da instituição e outros valores fixos, podem ser inseridos por meio de um redesenho do processo. Esse redesenho incluiria uma caixa de seleção com esses valores disponíveis no momento do cadastro, evitando a digitação manual e, consequentemente, erros. Isso representa uma melhoria no modelo de cadastro de dados.

- Para a dimensão de **consistência**, o redesenho de processo é a estratégia mais eficaz, pois trata da consistência entre dados e regras entre tabelas. Com base na análise do quadro 6.1, uma das inconsistência de consistência ocorre quando: a data de cadastro do ciclo dt_cadastro_ciclo é anterior à data de deferimento do curso dt_deferimento_curso, reformular o processo de cadastro de ciclos para incluir uma verificação lógica que garanta que a regra seja seguida. Sendo uma verificação no banco de dados ou no sistema que é utilizado para cadastrar os dados. Para as demais irregularidades de **consistência**, fazer também o redesenho.

7
Conclusão

Os dados do Sistec IFB estão disponíveis em formato CSV, conforme explicado na seção fonte de dados. A base de dados possui 120.873 registros, dos quais 80.521 apresentam pelo menos uma inconsistência, o que corresponde a 66,61% do total.

Para a análise da qualidade dos dados, inicialmente foram selecionadas 13 dimensões, apresentadas no capítulo na seção qualidade de dados. Contudo, devido a restrições no conhecimento detalhado sobre o gerenciamento e atualizações específicas do sistema, foi possível analisar apenas cinco dimensões objetivas: **unicidade, exatidão, completude, conformidade** e **consistência**. Essas cinco dimensões foram priorizadas por representarem aspectos fundamentais da confiabilidade dos dados.

Os resultados das métricas indicam que quatro dessas dimensões apresentaram valores superiores a 80%, considerados excelentes segundo MAKHOUL (2022), conforme apresentado no capítulo métrica global. Entretanto, a dimensão **consistência** apresentou um resultado de 33,60%, classificado como fraco, por estar entre 20% e 40%.

A dimensão **consistência** registrou um total de 80.252 inconsistências, representando 66,39% do total de registros (120.873). As demais dimensões apresentaram percentuais muito baixos de irregularidades: unicidade 0,02%, exatidão 0,20%, completude 0,0017% e conformidade 0,20%. Isso indica que as inconsistências relacionadas à **consistência** correspondem à grande maioria dos problemas, enquanto as outras dimensões somadas representam apenas 0,4217% das irregularidades.

Dos 80.521 registros que apresentam algum tipo de inconsistência, 80.252 estão relacionados à dimensão **consistência**, o que corresponde a 99,67% do total das inconsistências identificadas. Essa concentração evidencia que focar em ações específicas para corrigir e melhorar a dimensão **consistência** terá um impacto significativo na qualidade geral dos dados, reduzindo a quantidade total de inconsistências da base.

A grande quantidade de inconsistências na dimensão **consistência** pode acarretar problemas significativos, como relatórios imprecisos, decisões mal informadas e ineficiências nos processos institucionais. Por exemplo, datas de cadastro inconsistentes ou tipos de oferta divergentes podem causar falhas no acompanhamento acadêmico e na geração de relatórios confiáveis.

É importante destacar que este estudo teve limitações, principalmente devido ao acesso restrito a informações detalhadas do gerenciamento do sistema e a algumas colunas sensíveis de dados. Dessa forma, nem todas as dimensões de qualidade puderam ser avaliadas, o que sugere que pesquisas futuras poderiam oferecer uma visão mais abrangente da qualidade dos dados do Sistec IFB.

Ao analisar a qualidade dos dados por dimensão, os resultados da métrica global, apresentados na seção resultado da métrica global, indicam um valor aproximado de 86,62%. Essa classificação é considerada excelente, embora revele a existência de áreas críticas que demandam melhorias. A redução dessas inconsistências contribuirá para uma base de dados mais confiável, que exigirá menos tratamento durante a extração de informações, minimizando a perda de dados

importantes e aumentando a precisão das análises.

Aplicar as sugestões de melhorias apresentadas no capítulo Sugestões de melhorias para aumentar a qualidade dos dados elevará a qualidade dos dados e a precisão dos relatórios institucionais, apoiando melhores tomadas de decisão. Portanto, estabelecer processos contínuos de monitoramento da qualidade dos dados é fundamental para identificar e corrigir rapidamente novas inconsistências, garantindo que os avanços alcançados sejam mantidos ao longo do tempo e que a base de dados continue confiável.

7.1 Trabalhos futuros

Esta seção apresenta sugestões de trabalhos futuros que possibilitem a continuidade deste estudo, contribuindo para um aprofundamento nas análises da qualidade dos dados do Sistec IFB.

Sugestões de trabalhos futuros

As sugestões de trabalhos futuros são apresentadas a seguir em formato de lista:

- Analisar mais dimensões, caso no futuro seja possível obter mais informações sobre o gerenciamento, regras de segurança e atualização dos dados, para avaliar outras dimensões que não puderam ser analisadas devido às limitações no conhecimento dos dados.

- Avaliar a qualidade dos dados de outras instituições além do IFB, para realizar uma análise completa da qualidade dos dados de matrículas do Sistec em todo o Brasil, já que o Sistec armazena dados de matrículas de diversas instituições nacionais. Também, analisar todas as colunas existentes no Sistec, pois as colunas analisadas neste trabalho são aquelas com dados abertos. Caso seja possível analisar todas as colunas de todas as instituições, a análise da qualidade dos dados será mais ampla, permitindo identificar quais dimensões são afetadas e quais instituições apresentam maior ou menor qualidade dos dados.

- Aplicar técnicas de *Machine Learning*. Existem estudos que utilizam essa técnica para medir a qualidade dos dados, como o trabalho apresentado na revisão da literatura no capítulo Revisão da Literatura, intitulado *An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis* (ELOUATAOUI; EL MENDILI; GAHI, 2023).

Com essas sugestões, é possível dar continuidade ao trabalho, realizando análises mais aprofundadas e abrangentes. A aplicação de *Machine Learning* em dados já rotulados com inconsistências pode identificar irregularidades em novos dados, sem a necessidade de validações novamente nos dados utilizando as regras de negócio. Essas propostas de trabalhos futuros contribuem para aumentar as analises da qualidade dos dados do Sistec e uso de novas técnicas.

BATINI, C. et al. Methodologies for data quality assessment and improvement. **ACM computing surveys (CSUR)**, [S.l.], v.41, n.3, p.1–52, 2009.

BENTO, A. Como fazer uma revisão da literatura: considerações teóricas e práticas. **Revista JA (Associação Académica da Universidade da Madeira)**, [S.l.], v.7, n.65, p.42–44, 2012.

BRASIL. **Fala.BR - Plataforma Integrada de Ouvidoria e Acesso à Informação**. Acesso em: 26 maio. 2025.

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. **Data science journal**, [S.l.], v.14, p.2–2, 2015.

Conselho Nacional de Desenvolvimento Científico e Tecnológico. **Termo de Execução Descentralizada (TED)**. Disponível em: `https://www.gov.br/cnpq/pt-br/acesso-a-informacao/acoes-e-programas/parcerias/nacionais-1/termo-de-execucao-descentralizada-ted`. Acesso em: 21 out. 2024.

DANIEL, B. K. Reimaging research methodology as data science. **Big Data and Cognitive Computing**, [S.l.], v.2, n.1, p.4, 2018.

DONATO, H.; DONATO, M. Etapas na condução de uma revisão sistemática. **Acta Médica Portuguesa**, [S.l.], v.32, n.3, p.227–235, 2019.

ELOUATAOUI, W.; EL MENDILI, S.; GAHI, Y. An Automated Big Data Quality Anomaly Correction Framework Using Predictive Analysis. **Data**, [S.l.], v.8, p.182, 2023.

ELOUATAOUI, W. et al. An advanced big data quality framework based on weighted metrics. **Big Data and Cognitive Computing**, [S.l.], v.6, p.153, 2022.

FIGMA, I. **Figma**: ferramenta de design colaborativo para criar produtos significativos. acessado: 09/07/2025, `https://www.figma.com/pt-br/`.

Governo Digital. **Governança de Dados**. Disponível em: `https://www.gov.br/governodigital/pt-br/governanca-de-dados`. Acesso em: 21 out. 2024.

JUDDOO, S. et al. Data governance in the health industry: investigating data quality dimensions within a big data context. **Applied System Innovation**, [S.l.], v.1, n.4, p.43, 2018.

KRISHNA, C. M.; RUIKAR, K.; JHA, K. N. Determinants of data quality dimensions for assessing highway infrastructure data using semiotic framework. **Buildings**, [S.l.], v.13, n.4, p.944, 2023.

LIMA MACHADO, F. de. Censo Escolar e Sistec: as mais importantes bases de coleta de dados para ept. **Revista de Gestão e Avaliação Educacional**, [S.l.], p.1–8, 2019.

MAKHOUL, N. Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring. **Advances in Bridge Engineering**, [S.l.], v.3, n.1, p.17, 2022.

Ministério da Educação. **Pronatec**. Disponível em: http: //portal.mec.gov.br/component/content/article?id=34661:pronatec. Acesso em: 21 out. 2024.

Ministério da Educação. **Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec)**. Disponível em: http://portal.mec.gov.br/sistec-inicial. Acesso em: 21 out. 2024.

OLIVEIRA, F. H. M. et al. Painel de dados do Sistec: uma ferramenta para apoio à tomada de decisão. In: SIMPÓSIO BRASILEIRO SOBRE FATORES HUMANOS EM SISTEMAS COMPUTACIONAIS (IHC). **Anais...** [S.l.: s.n.], 2024. p.18–22.

PANIAN, Z. Some practical experiences in data governance. **World Academy of Science, Engineering and Technology**, [S.l.], v.62, n.1, p.939–946, 2010.

PYTHON Documentation. acessado: 09/07/2025, https://www.python.org/doc/.

SIMEC. **Termos de Execução Descentralizada**. Disponível em: https: //simec.mec.gov.br/ted/termo-de-execucao-descentralizada.php. Acesso em: 21 out. 2024.

TALEB, I. et al. Big data quality: a quality dimensions evaluation. In: INTL IEEE CONFERENCES ON UBIQUITOUS INTELLIGENCE & COMPUTING, ADVANCED AND TRUSTED COMPUTING, SCALABLE COMPUTING AND COMMUNICATIONS, CLOUD AND BIG DATA COMPUTING, INTERNET OF PEOPLE, AND SMART WORLD CONGRESS (UIC/ATC/SCALCOM/CBDCOM/IOP/SMARTWORLD), 2016. **Anais...** [S.l.: s.n.], 2016. p.759–765.

Tribunal de Contas da União. **O que é Governo Digital?** Disponível em: https: //portal.tcu.gov.br/fiscalizacao-de-tecnologia-da-informacao/atuacao/governo-digital/. Acesso em: 21 out. 2024.

ULARU, E. G. et al. Perspectives on big data and big data analytics. **Database Systems Journal**, [S.l.], v.3, n.4, p.3–14, 2012.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of management information systems**, [S.l.], v.12, n.4, p.5–33, 1996.

ZAKIR, J.; SEYMOUR, T.; BERG, K. Big Data analytics. **Issues in Information Systems**, [S.l.], v.16, n.2, p.81–90, 2015.

APÊNDICE A – Documentação dos Dados do Sistec IFB

Figura 7.1: Documentação do exemplo apresentado na Figura 4.6.

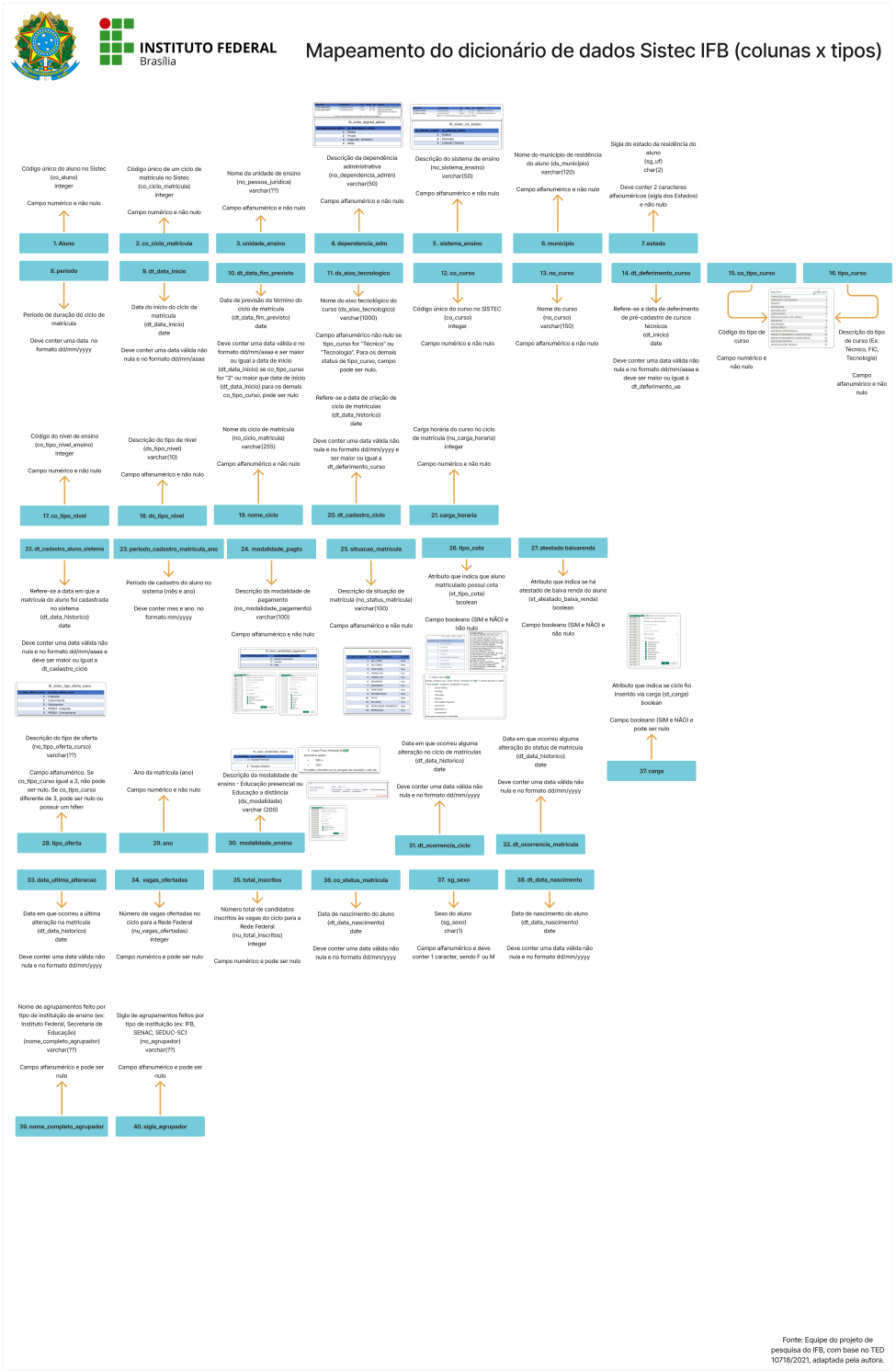


Figura 7.2: Documentação do exemplo apresentado na Figura 4.7.