PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Erickson Flávio Rosa

Qualidade dos Dados para Projetos de Aprendizado de Máquina - Revisão Sistemática de Literatura e Proposta de Métricas

Belo Horizonte

2025

Erickson Flávio Rosa

Qualidade dos Dados para Projetos de Aprendizado de Máquina - Revisão Sistemática de Literatura e Proposta de Métricas

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Dr. Luis Enrique Zárate

Belo Horizonte

2025

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Rosa, Erickson Flávio
R789q Qualidade dos dados para projetos de aprendizado de máquina: revisão sistemática de literatura e proposta de métrica / Erickson Flávio Rosa. Belo Horizonte, 2025.
79 f. : il.

Orientador: Luis Enrique Zárate

Dissertação (Mestrado) - Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Aprendizado do computador. 2. Ciência de dados. 3. Revisão Sistemática. 4. Controle de qualidade. 5. Mineração de dados (Computação). 6. Processamento de dados. I. Zárate, Luis Enrique. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.03

Ficha catalográfica elaborada por Fabiana Marques de Souza e Silva - CRB 6/2086

Erickson Flávio Rosa

Qualidade dos Dados para Projetos de Aprendizado de Máquina - Revisão Sistemática de Literatura e Proposta de Métricas

Dissertação apresentada ao Programa de Pós-Graduação em Informática como requisito parcial para qualificação ao Grau de Mestre em Informática pela Pontifícia Universidade Católica de Minas Gerais.

Prof. Dr. Luis Enrique Zárate - PUC Minas
(Orientador)

Prof.^{a} Dr.^{a} Cristiane Neri Nobre - PUC Minas (Banca Examinadora)

Prof. Dr. Carlos Santos Pires - UFCG
(Banca Examinadora)

Belo Horizonte, 26 de Agosto de 2025.

A Deus, minha mãe e minha família

AGRADECIMENTOS

Agradeço primeiramente a Deus, sem o qual não teria a força e o direcionamento necessários à conclusão desse programa.

Agradeço também à minha família, que foi penalizada em diversos momentos em que precisei priorizar o programa e, mesmo assim, não deixou de me apoiar com sua compreensão e complacência. E não poderia deixar de agradecer especialmente à Maria Madalena dos Santos, minha mãe de coração e de alma, que é seguramente minha maior referência de vida e responsável pelas maiores conquistas que tive, além de ter sido minha maior incentivadora nos estudos. Ela já se foi desse mundo, mas tenho certeza que de algum lugar ela me observa, e espero que esteja orgulhosa do seu legado, mostrando que sua missão aqui foi muito bem cumprida.

Agradeço a todos do Programa de Pós Graduação em Informática, que me possibilitaram muito aprendizado e evolução. Agradecimento especial ao meu orientador, Prof. Dr. Luis Enrique Zárate, que além de colocar à disposição todo seu conhecimento e sua experiência acadêmica, mostrou grande parceria e amizade durante todo nosso trabalho. E nessa grande jornada, que me trouxe diversos desafios e na qual várias vezes cheguei a fraquejar, sua postura de parceria e seu exemplo de firmeza certamente foram determinantes para minha permanência e dedicação na conclusão do programa.

Também preciso agradecer meus colegas de trabalho e amigos da Prodabel, especialmente meu gerente João Augusto, que na reta final dessa jornada fez questão de reforçar a importância dessa conquista para meu crescimento pessoal e profissional e de me dar todo apoio que podia para concluí-la com sucesso. Também preciso destacar o agradecimento à minha colega de trabalho e amiga Bruna Siqueira, que me deu um apoio muito valioso em um momento crítico desse programa.

Por fim, agradeço a todos os grandes amigos que contribuíram de alguma forma com toda essa jornada, ainda que com palavras de incentivo e orgulho. Dentre esses, destaco a Profa. Aldeny Santos, que contribuiu significativamente para nascer em mim a sede pelo conhecimento e o gosto pelo estudo, e o amigo Davidson Rocha, cuja força demonstrada ao lidar com o enorme desafio que se colocou diante dele durante essa minha jornada fez com que eu passasse a tê-lo também como referência de garra e superação.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."

Albert Einstein

RESUMO

Com o crescimento exponencial do volume de dados gerados pelas organizações, cresce também o interesse delas por soluções que lhes permitam tirar o máximo proveito desses dados. Nesse cenário destaca-se a Ciência de Dados, que possui técnicas e ferramentas desenvolvidas justamente para esse propósito. Mas a qualidade dos resultados de um Projeto de Ciência de Dados é extremamente dependente da qualidade dos dados que serão utilizados como entrada no processo. Na área da Ciência de Dados, quando nos referimos à qualidade do conjunto de dados para construção de modelos de aprendizado, existem diversos aspectos que podem afetar a qualidade dos modelos resultantes. Neste trabalho é apresentada uma revisão sistemática da literatura que buscou identificar os principais aspectos a serem considerados para se avaliar a qualidade do conjunto de dados a serem utilizados em projetos de Aprendizado de Máquina, que é uma das principais áreas da Ciência de Dados. A revisão mostra que o tema tem sido ainda pouco explorado e, pela falta de métricas apresentadas na literatura, neste trabalho é proposta uma metodologia para a avaliação da qualidade dos dados para projetos de Aprendizado de Máquina, composto por 12 dos aspectos apontados pela literatura e contendo também uma proposta de métricas de qualidade para eles.

Palavras-chave: ciência de dados, aprendizado de máquina, aspectos de qualidade de dados, métricas de qualidade de dados

ABSTRACT

With the exponential growth in the volume of data generated by organizations, their interest in solutions that allow them to make the most of this data is also growing. In this scenario, Data Science stands out, with techniques and tools developed precisely for this purpose. However, the quality of the results of a Data Science Project is extremely dependent on the quality of the data that will be used as input in the process. In the area of Data Science, when we refer to the quality of the data set for building learning models, there are several aspects that can affect the quality of the resulting models. This paper presents a systematic review of the literature that sought to identify the main aspects to be considered when evaluating the quality of the data set to be used in Machine Learning projects, which is one of the main areas of Data Science. The review shows that the topic has still been little explored and, due to the lack of metrics presented in the literature, this work proposes a data quality assessment model for Machine Learning projects composed of 12 of the aspects highlighted in the literature, also containing a proposal for quality metrics for them.

Keywords: data science, machine learning, data quality aspects, data quality metrics

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

ACM – Association for Computing Machinery

DLC – Data Lifecycle

DS – Dataset

FU – Função Utilidade

FUT – Função Utilidade Total

IEEE – Institute of Electrical and Electronics Engineers

LGPD – Lei Geral de Proteção de Dados Pessoais

RA – Research Answers

RQ – Research Questions

RS – Research String

SLR – Systematic Literature Review

UCI – University of California, Irvine

SUMÁRIO

1 INTRODUÇÃO

A adoção massiva de sistemas informatizados, nos mais diversos setores das organizações, tem gerado um vertiginoso aumento na coleta e armazenamento de dados sobre os mais diversos processos relacionados às suas atividades operacionais e de serviços. À medida em que as organizações e o público que almejam atingir aumentam suas demandas, aumenta também o interesse em utilizar estrategicamente esses dados. Por meio de modelos de aprendizado computacional, as organizações podem compreender e gerenciar mais profundamente seus processos e melhorar seus resultados. Esse cenário tem estimulado nas organizações a necessidade por adquirir cada vez mais conhecimento sobre seus dados, permitindo extrair *insights* e informações relevantes para melhorar as tomadas de decisões, além de obter vantagens competitivas.

De forma a obter maior assertividade nas tomadas de decisões, os modelos devem ser confiáveis e representativos, e isto começa com a qualidade dos dados sobre diversos aspectos. Por exemplo, quando nos referimos à área da saúde, o conhecimento a ser adquirido por meio dos modelos de aprendizado de máquina demanda mais garantia da representatividade e assertividade. Na área social as maiores demandas são pela ética e pela justiça. Esses diversos aspectos, quando levados em consideração, contribuem no aperfeiçoamento dos sistemas de diagnósticos, aumentando a eficácia das ações considerando os fins para os quais foram construídos.

Nesse cenário de demandas, a área da Ciência de Dados visa adquirir conhecimento e obter valor a partir dos dados gerados nos mais diversos setores e processos. Entre os produtos gerados por esta área tem-se: análise de históricos; construção de modelos de aprendizado descritivos, preditivos ou prescritivos; identificação de novas hipóteses; etc. A área também trata da geração de repositórios de dados para diversos outros fins, analisando a disponibilidade dos dados, a integração de bases de dados, dentre outros artefatos (PENG; MATSUI, 2018). Tudo isso permitindo uma visão mais abrangente e assertiva para tomada de decisão a partir de fatos ocorridos que possam ser expressos por meio de dados.

A eficácia e até mesmo a utilidade dos diversos artefatos gerados pelo Aprendizado de Máquina dependem profundamente da qualidade dos dados que são utilizados como insumo. Como qualidade, não se deve considerar apenas os aspectos técnicos e de infraestrutura, mas também os semânticos, que apontam para a capacidade que o conjunto de dados possui de entregar um resultado alinhado ao propósito do projeto. O projeto pode

utilizar as melhores ferramentas, as melhores técnicas, os melhores algoritmos e ser realizado pelos melhores profissionais, mas ainda assim não fornecer resultados satisfatórios por não se basear em dados que atendam aos aspectos aplicáveis de qualidade.

A partir do exposto, surgem então algumas questões que devem ser respondidas: Quais são as principais contribuições científicas para a avaliação da qualidade dos dados para projetos de Aprendizado de Máquina? Quais são as metodologias e métricas para avaliar essa qualidade? É na busca por respostas a essas questões que este trabalho se insere. Para tal, foi realizada um revisão sistêmica da literatura que teve como produto resultante a identificação dos aspectos considerados relevantes para avaliar a qualidade dos conjuntos de dados que serão utilizados em projetos de Aprendizado de Máquina. A partir dessa revisão foi possível observar a carência de métricas para tal fim, pois nenhuma foi proposta nos trabalhos encontrados.

Tendo como objetivo a construção de modelos de aprendizado adequados ao domínio de problema, alguns aspectos de qualidade podem ter maior ou menor utilidade do que outros aspectos. Para isso, é associada a cada métrica uma Função Utilidade (FU), que pode ser ajustada de acordo com os objetivos que se deseja alcançar no projeto. Desta forma, uma Função Utilidade Total (FUT), de somas ponderadas, pode ser calculada considerando as funções de utilidade parciais. O valor da FUT pode refletir uma nota final para o conjunto de dados a ser utilizado na construção do modelo de aprendizado. A proposta é avaliada a partir de sua aplicação em um conjunto de dados Toy e também em bases de dados reais.

1.1 Referencial Teórico

No campo da engenharia de software existe a norma ISO25010, de 2011, que define um modelo para elevar os produtos de *software* desenvolvidos a altos níveis de qualidade. Esse modelo aponta diversas etapas que devem ser consideradas durante o desenvolvimento de softwares, inclusive definindo aspectos mais precisos a serem avaliados durante todo o processo de desenvolvimento, sempre buscando a melhoria contínua do produto final resultante (ISO/IEC 25010, 2011). O Aprendizado de Máquina pode ser considerada uma área emergente e, mesmo tendo recebido nos últimos anos uma grande atenção de pesquisas e desenvolvimentos, tanto pela comunidade acadêmica quanto pela profissional, talvez ainda não tenha a mesma maturidade no estudo e investigação de todos os temas pertinentes à área, como a pesquisa na qualidade dos dados e sua relação com os objetivos de um projeto para descoberta de conhecimento.

Quando refere-se à qualidade do conjunto de dados para a construção de modelos de aprendizado e extração de padrões, existem diversos aspectos que podem afetar a qua-

lidade dos modelos. Em trabalhos iniciais, como (BATINI et al., 2009), são apontados quatro aspectos como referência para a qualidade dos dados: acurácia, completude, consistência e temporalidade. Em (ALAOUI; GAHI; MESSOUSSI, 2019), os autores definem um conjunto de aspectos para avaliar a qualidade do conjunto de dados que será utilizado em projetos de *big data* para análise de sentimentos. Esses aspectos são organizados em três grupos: confiabilidade, usabilidade e pertinência. Também definiram um conjunto de métricas para cada um desses aspectos. No aspecto confiabilidade os autores propõem as seguintes medidas: acurácia, completude e unicidade. No aspecto usabilidade, as medidas: transformação, conformidade, penalidade de armazenamento, normalização e integridade referencial. Já em pertinência, as medidas: consistência, credibilidade e frescor. Diversos são os termos e as definições dados aos aspectos de qualidade na literatura.

Em (HE et al., 2019), os autores consideram quatro aspectos para definir a qualidade dos dados em projetos de Aprendizado de Máquina para *deep learning*, que são: distorção das classes, complexidade da amostra, qualidade dos rótulos e dados ruidosos. No trabalho apresentado por Rudraraju e Boyanapally (RUDRARAJU; BOYANAPALLY, 2019) foi realizada uma entrevista com 15 cientistas de dados para levantar aspectos relevantes na avaliação da qualidade de dados em projetos de Aprendizado de Máquina. Ao final desse trabalho foram apontados 16 aspectos, sendo eles: precisão, integridade, consistência, volatilidade, relevância, interpretabilidade, eficácia, eficiência, satisfação, contexto de cobertura, liberdade de risco, privacidade, reprodutibilidade, tamanho, diversidade e justiça.

Em (ARASS; TIKITO; SOUISSI, 2017), os autores realizam uma revisão acerca das principais *frameworks* propostas na literatura para definir o Data Lifecycle (DLC). Resumidamente, esses *frameworks* consideram as fases de *data creation, data processing and storage, data usage, data archiving and data destruction.* Em (ARASS; SOUISSI, 2018), os autores propõem um novo DLC chamado Smart DLC para auxiliar na transformação de dados brutos em *smart data*, para contextos de *big data.* O *framework* proposto considera 14 fases no ciclo de vida: Planejamento, Gerenciamento, Coleta, Integração, Filtragem, Enriquecimento, Análise, Acesso, Visualização, Armazenamento, Destruição, Arquivamento, Segurança e Qualidade. Os autores ainda ressaltam diversas visões que a literatura tem dado para o aspecto qualidade.

Em (CAI; ZHU, 2015), os autores definem aspectos a serem observados na avaliação da qualidade dos dados, mais especificamente em projetos de *big data.* Os aspectos apontados pelos autores são: acessibilidade, temporalidade, autorização, credibilidade, definição/documentação, metadados, precisão, consistência, integridade, completude, auditabilidade, capacidade, legibilidade e estrutura.

Para alguns autores o controle de qualidade deve ocorrer durante as transições de

uma fase para outra do ciclo de vida dos dados. Isso seria feito por meio da definição de requisitos de qualidade, do nível de precisão esperado, e da proposta de medidas para avaliar a satisfação com a qualidade dos dados. Para outros autores, o controle da qualidade é alcançado introduzindo protocolos que devem ser aplicados para garantir que os dados sejam devidamente recolhidos, geridos, processados, utilizados e mantidos em todas as fases do seu ciclo de vida. No trabalho de (ARASS; SOUISSI, 2018) é proposto que se avalie duas fases do DLC para a construção de modelos de aprendizado: Acesso e Análise.

1.2 Problema

Diversos projetos de Aprendizado de Máquina não conseguem resultados com a qualidade desejada e isso geralmente só é percebido ao final do projeto. A avaliação prévia de alguns fatores no início do projeto, como a qualidade dos dados que serão utilizados pelo projeto para realizar o aprendizado do modelo, pode evitar ou ao menos reduzir os riscos desse tipo de ocorrência. Nesse contexto, o problema a ser tratado nesta dissertação é a carência de formas de avaliação da qualidade destes dados.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é a identificação de aspectos de qualidade de dados, a elaboração de uma proposta de métricas e a definição de funções de utilidade que possam ser aplicadas para avaliar conjuntos de dados utilizados em projetos de Aprendizado de Máquina.

1.3.2 Objetivos Específicos

A fim de alcançar os objetivos gerais apontados acima, foram estabelecidos os seguintes objetivos específicos:

a) Realizar uma Revisão Sistemática da Literatura (Systematic Literature Review (SLR)) para identificar os aspectos de qualidade apontados como sendo os que devem ser avaliados em *datasets* que serão utilizados em projetos de Ciência de Dados e Aprendizado de Máquina.

b) Elaborar uma proposta de métricas de avaliação para um conjunto de aspectos selecionados dentre os apresentados na Revisão Sistemática da Literatura do item

anterior, dentro do conceito de DLC (ARASS; SOUISSI, 2018), correspondentes ao Acesso e à Análise do conjunto de dados. Esta proposta deverá conter métricas e também Funções de Utilidade que permitirão ponderar os resultados de acordo com o esperado para o modelo de aprendizado que será gerado. A ideia é que, quanto maior for o resultado da Função de Utilidade, mais o conjunto de dados contribuirá para melhorar o modelo de aprendizado.

c) Aplicar a proposta de métricas em um *dataset toy* para exemplificar seu uso de forma didática. Aplicar também a proposta em *datasets* reais e apresentar uma discussão dos resultados.

e) Realizar experimentos com a aplicação da metodologia proposta a fim de avaliar sua assertividade através da utilização de um algoritmo de Aprendizado de Máquina.

1.4 Justificativa

A utilização de uma metodologia de avaliação da qualidade dos dados já na fase inicial do projeto possibilita a identificação de necessidade de melhoria dos dados ou mesmo da obtenção de mais dados de forma a aumentar a chance de sucesso do projeto. Então, a intenção é que no final desse trabalho seja apresentada uma proposta de métricas de avaliação da qualidade desses dados cuja aplicação possa trazer melhoras significativas na qualidade do Aprendizado de Máquina resultante do projeto, redução do seu custo e tempo de execução ou até mesmo todos esses benefícios.

1.5 Organização

A organização deste trabalho será da seguinte forma: no Capítulo 2 será apresentada uma Revisão Sistemática da Literatura realizada para encontrar os aspectos apontados como importantes na avaliação da qualidade dos dados para projetos de Ciência de Dados e Aprendizado de Máquina e também identificar a existência de propostas de modelos de avaliação da qualidade; no Capítulo 3 será apresentada a proposta de métricas para avaliação da qualidade em 12 aspectos selecionados dentre os apontados pela literatura, incluindo o cálculo de métricas de qualidade e funções utilidade; no Capítulo 4 será demonstrada a aplicação da proposta de métricas em um conjunto de dados *toy*; no Capítulo 5 serão exibidos os resultados da aplicação da proposta de métricas em 5 bases de dados selecionadas no UCI e alguns experimentos com essas bases para verificação da assertividade da proposta de métricas; por fim, o Capítulo 6 trará as conclusões do trabalho realizado e algumas possibilidades de estudos futuros para evolução desse trabalho.

2 REVISÃO SISTEMÁTICA DA LITERATURA

A revisão da literatura foi realizada aplicando a metodologia proposta em (KITCHENHAM et al., 2009), que organiza o trabalho em três fases: definição das questões de pesquisa, processo de busca, definição e aplicação dos critérios de inclusão e exclusão. A seguir, serão apresentadas as execuções destas fases para obtenção do material bibliográfico desejado.

2.1 Questões de Pesquisa

As questões de pesquisa a serem respondidas são:

- RQ1: Quais aspectos de qualidade devem ser considerados para avaliar o conjunto de dados a ser aplicado em projetos de Ciência de Dados?
- RQ2: Quais são as métricas propostas para avaliação da qualidade dos dados a serem aplicadas em projetos de Ciência de Dados?

Para esta revisão, foram também considerados trabalhos que apenas discorram acerca dos aspectos de qualidade, apontando-os e justificando sua importância, mesmo que não apresentem uma medida ou métrica para sua avaliação.

2.2 String de Busca

A seguir é apresentada a *string-de-busca* (Research String (RS)) que foi definida a fim de recuperar todos os trabalhos que possam responder às questões de pesquisa:

- RS: (“data science” OR “machine learning” OR “data mining” OR “artificial intelligence”) AND (“data quality” OR “quality attributes” OR “quality aspects” OR “ethical”)

Na primeira parte da conjunção da *string-de-busca* são considerados o principal termo “data science” e outros termos que são englobados por esse. Na segunda parte da *string-de-busca* são considerados os termos cuja aparição na publicação pode sinalizar que a mesma trata do assunto de interesse. Vale ressaltar que o termo “ethical” foi utilizado

porque as questões éticas recentemente têm chamado a atenção da comunidade de Ciência de Dados.

Mesmo com a junção dessas duas partes, a *string-de-busca* é bastante abrangente. A *string* foi montada dessa forma para diminuir o risco de alguma publicação relevante ao estudo não ser exibida no processo de busca. A busca nos repositórios ficou restrita ao título, resumo e palavras-chaves, dentro do período de 1996 e 2022. Como o termo "data mining" historicamente surgiu apenas em 1996, considerou-se que qualquer trabalho publicado antes desse ano não atenderia especificamente ao objetivo ou traria uma contribuição irrelevante, considerando as evoluções ocorridas na área após essa data.

2.3 Repositórios Pesquisados

Foram considerados os três principais repositórios de publicações na área: Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE) e Science Direct. Além desses, foi considerado também o repositório de publicações da área da saúde Pubmed. Esse repositório além de ter recebido atenção da comunidade de Ciência de Dados, corresponde a uma área onde a qualidade dos dados é um fator relevante. Na segunda coluna da Tabela 1 é mostrado o tamanho do *corpus* do material bibliográfico alcançado pelo processo de busca.

2.4 Critérios de Exclusão

Após o processo de buscas nos repositórios, filtros de exclusão foram executados conforme detalhado a seguir:

- **EC1 – Filtragem pelo tipo de publicação:** foi aplicado a todas as publicações retornadas pela busca. Foram mantidos apenas artigos científicos, dissertações e teses acadêmicas publicados em idioma inglês. Após a sua aplicação, das 853 publicações encontradas na busca, restaram 341 referências.
- **EC2 - Remoção por duplicação:** após aplicação da etapa EC1, publicações retornadas mais de uma vez pela busca, no mesmo repositório ou em repositórios distintos foram excluídas. Foram identificadas 3 ocorrências, atualizando o corpus bibliográfico para 338 publicações.
- **EC3 - Filtragem pelo título:** foi aplicado a todas as publicações resultantes após aplicação da etapa EC2. Por meio da leitura do título, foi aplicado o julgamento humano de sua relação direta com o objetivo desse trabalho. Após esse processo resultaram 124 publicações.

- **EC4 - Filtragem pelo resumo:** foi aplicado a todos os trabalhos restantes após a aplicação da EC3. Foi realizada a leitura do resumo e aplicado o julgamento humano de sua relevância com o objetivo desse trabalho, restando após essa etapa 77 publicações.

- **EC5 - Filtragem pela leitura diagonal:** foi aplicado a todos os trabalhos restantes após a aplicação da EC4, consistindo de uma leitura diagonal para levantamento da relevância de cada publicação ao objetivo desse trabalho. Após essa etapa restaram 42 publicações.

- **EC6 - Filtragem pelo leitura do conteúdo:** foi aplicado a todos os trabalhos restantes após a aplicação do EC5, durante a leitura integral do conteúdo de cada contribuição. No final, o *corpus* bibliográfico resultou em 12 publicações.

As quantidades de publicações resultantes de cada fase da pesquisa, por repositório, incluindo a busca, a aplicação dos critérios de exclusão e também os totais finais, podem ser visualizadas na Tabela 1. A vinculação dos trabalhos com os aspectos apontados por cada um deles pode ser vista na Tabela 2.

Tabela 1 – Resultados das buscas e dos critérios de exclusão

		Filtrado por					
Repositório	*Busca*	*EC1*	*EC2*	*EC3*	*EC4*	*EC5*	*EC6*
ACM	348	63	63	25	16	10	7
IEEE	191	16	16	7	3	2	0
Pubmed	101	64	63	60	46	23	3
Science Direct	213	198	196	32	12	7	2
Totais	853	341	338	124	77	42	12

Fonte: Dados da pesquisa

2.5 Análise das questões de pesquisa

Nesta seção, serão respondidas (Research Answers (RA)) as Questões de Pesquisa (Research Questions (RQ)), ressaltando os principais aspectos de qualidade apontados pela literatura, assim como também será realizada uma discussão acerca de medidas e métricas para avaliação da qualidade dos dados em projetos de Aprendizado de Máquina.

RA1: Aspectos de Qualidade a partir da Literatura. Nesta seção será respondida a questão RQ1. A seguir são listados, juntos com suas definições, os principais aspectos de qualidade mencionados pelos trabalhos resultantes da revisão da literatura. Os quatro (4) primeiros são referentes à fase de Acesso do DLC, e os oito (8) restantes à

fase de Análise. A relação de aspectos com trabalhos correspondentes ser vista na Tabela 2.

1) Rastreabilidade: avalia se é possível rastrear os dados em todas as etapas de sua obtenção, começando da fonte de origem até o dado disponibilizado para compor o *dataset*. Esse aspecto é bastante importante, principalmente quando existem dados extraídos de fontes externas ao projeto, como fornecedores externos, nuvens e *datacenters* instalados em outras localidades. Como exemplo, considere um conjunto de dados contendo dados demográficos e meteorológicos que é disponibilizado em nuvem pública e acessado através de portal de um órgão público. Esse conjunto de dados pode passar por diversos processos de extração, movimentação e carga de dados até ser disponibilizado para consumo em um projeto de Ciência de Dados. Na métrica proposta neste trabalho, um atributo do conjunto de dados será considerado rastreável se em todas as etapas dos processos de extração, movimentação e carga, o seu valor puder ser alcançado e inspecionado.

2) Integridade: avalia as inconsistências e erros presentes nos dados que podem ser causados por falhas de concepção, implementação ou execução nos processos de extração e transformação que geram os dados para o conjunto de dados. Como exemplo, considere a extração de um atributo numérico de altíssima precisão e que, por limitação da ferramenta que extrai e disponibiliza esses dados no conjunto de dados, o valor desse atributo é truncado ou arredondado, ficando com um valor diferente do dado original. Para verificar se os dados foram mantidos íntegros pode ser gerado um *hash* do dado original e um *hash* do dado disponibilizado no conjunto de dados. Em seguida esses *hashes* são comparados e, se eles forem iguais, significa que a integridade do dado foi mantida.

3) Privacidade: avalia se os dados possuem restrições quanto à sua exposição nas diversas etapas de um projeto de Ciência de Dados. Essas restrições podem ser necessárias tanto para evitar o uso indevido desses dados, como na execução de fraudes ou outros fins, quanto também para evitar a exposição da pessoa envolvida. Além disso, com a crescente disseminação e cobrança legal das leis de proteção de dados pelo mundo, esse aspecto tem se tornado um dos mais importantes ao se avaliar a concepção de projetos baseados em dados.

Como exemplo de dados que hoje são considerados sensíveis à privacidade pode-se citar o número do documento de identificação do cidadão, que não pode ser exposto em qualquer lugar sem mascaramento, principalmente para evitar o risco de fraudes. Quando algum atributo recebe, por exemplo, um índice alto para esse aspecto, isto significa que o atributo não é sensível à privacidade. Caso receba um índice baixo, o atributo é sensível à privacidade e pode ter uma significativa redução da sua contribuição no projeto, dificultar

a interpretabilidade dos resultados do modelo de aprendizado, ou até mesmo ter a perda de sentido da sua utilização.

4) Disponibilidade: avalia se o atributo está disponível sempre que uma nova extração de dados for necessária. Portanto, esse aspecto de qualidade só é aplicável a projetos cujos dados possuem volatilidade que extrapola a vida útil do modelo de aprendizado e, com isso, demandam processos periódicos de atualização desses dados. É importante ressaltar que esse aspecto considera a acessibilidade aos dados voláteis, já nos formatos e locais definidos, apenas nas datas e horários previamente estipulados no projeto para sua aquisição. Por exemplo, se for definido em um projeto de Ciência de Dados que a atualização dos dados deve ocorrer diariamente às 00:00 horas, no formato csv, e em determinado servidor de compartilhamento, somente esses requisitos precisarão ser atendidos para que esse aspecto seja considerado como satisfeito.

5) Relevância: avalia a importância dos atributos para as análises a serem feitas. Corresponde à relação do atributo com o domínio de estudo ou, mais precisamente, ao grau de influência do atributo nas inferências que serão realizadas pelo modelo de aprendizado. Um atributo X é relevante, se conhecido seu valor, pode alterar a estimação do rótulo de uma classe Y, ou seja, se Y é condicionalmente dependente de X. Em (KOHAVI; JOHN, 1997), dois níveis de relevância *Fraco* e *Forte* foram estabelecidas. Por exemplo, o consumo de sal em excesso, assim como a falta de exercícios físicos, estão fortemente associados com a hipertensão. A idade pode ser considerada fracamente associada com a doença.

6) Interpretabilidade: avalia se é possível compreender, em análise humana, a informação contida em todos os valores do atributo. Essa compreensão pode facilitar a percepção de valores incorretos ou discrepantes que poderiam comprometer a qualidade dos resultados do modelo de aprendizado. Esse aspecto pode facilitar o entendimento, interpretação e a influência dos valores de determinado atributo, nos resultados alcançados pelos modelos de aprendizado. Como exemplo, considere-se um atributo contendo o grupo sanguíneo de pacientes, onde o domínio de valores possíveis é pequeno e quem está realizando o estudo tem pleno conhecimento do significado dos valores desse domínio. Nesse caso, a medida desse aspecto deveria ser alta.

7) Consistência: avalia a coerência dos valores dos atributos em relação às instâncias do mundo real no qual eles estão inseridos. Esse aspecto deve ser avaliado nas dimensões sintática e semântica. Na avaliação sintática valida-se cada atributo, comparando seu valor com a definição do domínio correspondente. Na avaliação semântica valida-se a instância, verificando se a combinação dos valores dos seus atributos existe no mundo real, ou seja, no domínio representado. Como exemplo considere-se o atributo idade para uma pessoa em anos contendo o valor de 4 anos. Esse valor é considerado sintaticamente

correto porque pertence ao domínio de valores do atributo, uma vez que existem no mundo real pessoas com essa idade. Porém, será considerado semanticamente incorreto se em outro atributo da mesma classe houver a informação de que possui filhos.

8) Diversidade: determina se o conjunto de dados possui quantidade significativa de instâncias distintas considerando todo o espaço do universo possível. A não existência de instâncias suficientes pode indicar que os modelos de aprendizado passem a ter um viés para as instâncias representadas e perda de capacidade de generalização pela falta de instâncias suficientes para representar o espaço do universo. Esta medida proporciona uma ideia do tamanho da amostra considerada na construção de um modelo de Aprendizado de Máquina. É importante notar que esta medida não impede a construção de modelos. Porém, é um indicador da representatividade do modelo para um domínio de problema, e das restrições que podem ser impostas aos modelos construídos. Como exemplo, considere-se a fabricação de veículos autônomos e o conhecimento dos tipos de defeitos que todos seus componentes podem apresentar. Contendo um conjunto de dados dos veículos que apresentaram defeito, e sendo o objetivo identificar padrões de defeitos para definir procedimentos de reparo, provavelmente não seja possível cobrir todo o espaço de possibilidades de defeitos. Daí os procedimentos de detecção de falhas e reparos podem não ser genéricos o suficiente.

9) Eficácia: tem o propósito primário de fornecer uma expectativa na obtenção de novo conhecimento, sendo útil e relevante. Para isso, é calculada a diferença da quantidade de informação contida em um conjunto de dados em relação ao conjunto de dados de referência (o qual contém todas as instâncias possíveis para o domínio sendo representado). Quanto menor a diferença na quantidade de informação, maior será a eficácia para obtenção de novo conhecimento. Em (UNGER; HARN; KUMAR, 1990) é proposto quantificar a informação contida no conjunto de dados, bastando considerar cada registro como uma mensagem única, eliminando previamente as redundâncias. Como exemplo, num estudo que se propõe diagnosticar uma doença utilizando resultados de exames clínicos diversos, quanto mais informações tivermos nesse conjunto de dados, maior será a capacidade de diagnosticar corretamente a doença.

10) Justiça: tenta alertar previamente que o conhecimento extraído do conjunto de dados, por meio do modelo de aprendizado, pode apresentar comportamento discriminatório ou de classificação igualitária para os desiguais. Nesse aspecto analisa-se apenas os atributos sensíveis às questões de justiça, que podem ser aqueles vinculados às definições de etnia, gênero, faixa etária, condição social e nível educacional. Para avaliar esse aspecto é possível considerar a distribuição uniforme dos valores de cada atributo. Como exemplo, considere-se uma análise de criminalidade feita sobre uma base de dados contendo crimes cometidos, onde a grande maioria dos autores é de certa etnia. Nessa

situação, o modelo de aprendizado poderia considerar que aquela etnia está mais propensa a praticar crimes. Um outro exemplo seria um estudo dos níveis educacionais de adolescentes segundo as condições sócio econômicas das suas famílias. Se o estudo fosse realizado em apenas um segmento da região geográfica estudada, onde a população é economicamente mais abastada, ou mais carente, poderiam ser incorporadas características discriminatórias no modelo de aprendizado.

11) Representatividade: determina se os dados de um atributo, disponibilizado no conjunto de dados (amostra), estão na mesma proporção da população que se deseja estudar. Essa avaliação pode considerar atributos que são relevantes para o estudo a ser realizado. Como exemplo, para estudo de uma população que possui a mesma proporção de homens e mulheres, a amostra deveria manter a mesma proporção. Quanto mais equilibrada a quantidade de homens e mulheres do conjunto de dados com a população, maior a representatividade do conjunto de dados e mais alto seria o índice desse aspecto.

12) Reprodutibilidade: avalia a expectativa do modelo de poder ser operado sob instâncias distintas da população. Para isso, é verificado se um determinado atributo, ou conjunto de atributos, se distribui da mesma forma em várias amostras de uma ou várias populações. Isso traria maior confiabilidade ao modelo de aprendizado. Como exemplo de aplicação desse aspecto, é possível considerar a necessidade latente de confirmar se para diversas instâncias de uma população, não observadas durante o processo de treinamento, o modelo apresentaria resultados confiáveis.

RA2: Métricas para avaliação da qualidade do conjunto de dados. A partir dos resultados da revisão do *corpus* da literatura considerado, não foi possível identificar métricas para aspectos de qualidade de dados para projetos de Aprendizado de Máquina. Por esse motivo, neste trabalho são propostas métricas para 12 dos aspectos de qualidade apresentados, sendo eles: Rastreabilidade, Integridade, Privacidade, Disponibilidade, Relevância, Interpretabilidade, Consistência, Diversidade, Eficácia, Justiça, Representatividade e Reprodutibilidade.

Tabela 2 – Publicações e aspectos de qualidade apontados por cada uma delas

	Levantamento		Contribuição			Aspectos											
Publicação	**Questionário**	**Revisão de Literatura**	**Modelo**	**Metodologia**	**Métrica**	**Rastreabilidade**	**Integridade**	**Privacidade**	**Disponibilidade**	**Relevância**	**Interpretabilidade**	**Consistência**	**Diversidade**	**Eficácia**	**Justiça**	**Representatividade**	**Reprodutibilidade**
P01 (DING et al., 2019)	X		X														
P02 (THIRUMURUGANATHAN et al., 2021)		X										X					
P03 (SCHELTER et al., 2018)	X		X	X							X	X					
P04 (BERTOSSI; GEERTS, 2020)	X							X							X		
P05 (BOYD, 2021)	X		X														
P06 (MICELI; POSADA; YANG, 2022)		X	X									X					
P07 (BLAKE; MANGIAMELI, 2011)		X															
P08 (CHEN; CHEN; DING, 2021)	X											X					
P09 (CHATILA et al., 2017)		X									X	X		X			
P10 (JIAN, 2019)		X				X	X		X			X					
P11 (BRADY; NERI, 2020)		X								X				X		X	
P12 (MARTINEZ-MARTIN et al., 2018)		X	X										X				
P13 (WINFIELD; JIROTKA, 2018)	X								X				X		X		
P14 (LAM et al., 2021)		X						X					X				
P15 (VAYENA; BLASIMME; COHEN, 2018)	X												X				
P16 (CARTER et al., 2020)		X								X				X	X		
P17 (CHEN et al., 2021)	X						X						X				
P18 (CHAR; ABRàMOFF; FEUDTNER, 2020)	X										X		X				
P19 (LANDAU et al., 2022)		X						X					X				
P20 (MCCRADDEN et al., 2020)	X												X				
P21 (MöLLMANN; MIRBABAIE; STIEGLITZ, 2021)		X									X	X		X			
P22 (PESAPANE et al., 2018)		X								X	X			X			
P23 (CURRIE; HAWK, 2021)		X							X						X		
P24 (BEIL et al., 2019)		X				X	X	X					X				
P25 (MULLINS; HOLLAND; CUNNEEN, 2021)	X												X				
P26 (ROSEMANN; ZHANG, 2022)	X						X						X				
P27 (SAHEB; SAHEB; CARPENTER, 2021)		X											X				
P28 (CLARKE, 2019)		X												X	X		X
P29 (RUDRARAJU; BOYANAPALLY, 2019)	X		X					X			X	X	X	X	X		

Fonte: Dados da pesquisa

3 PROPOSTA DE MÉTRICAS PARA OS ASPECTOS DE QUALIDADE

Neste capítulo, são propostas 12 métricas para avaliar os aspectos de qualidade apontados no capítulo anterior. As métricas atribuem um valor ao índice de qualidade para cada aspecto considerado, e são aplicadas somente a atributos categóricos. Sendo assim, atributos contínuos, caso existam, deverão ser categorizados antes da aplicação destas métricas.

Para a proposta das métricas são definidas a matriz do conjunto de dados e as variáveis, globais e específicas, utilizadas para os cálculos dos índices de qualidade.

3.1 Variáveis Globais

A	$\Longrightarrow$	Representa o conjunto de dados contendo instâncias únicas
M	$\Longrightarrow$	Quantidade de instâncias do conjunto de dados
N	$\Longrightarrow$	Quantidade de atributos do conjunto de dados
m	$\Longrightarrow$	Índice para a instância no conjunto de dados, onde $m = 1..M$
n	$\Longrightarrow$	Índice para o atributo no conjunto de dados, onde $n = 1..N$
a_m	$\Longrightarrow$	Valores de todos os atributos da m-ésima instância de A
a_n	$\Longrightarrow$	Valores do n-ésimo atributo em todas as instâncias de A
a_{mn}	$\Longrightarrow$	Valor do atributo n na instância m
NA	$\Longrightarrow$	Quantidade de aspectos avaliados no conjunto de dados

3.2 Matriz do Conjunto de Dados

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2N} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3N} \\ \dots & \dots & \dots & \dots & \dots \\ a_{M1} & a_{M2} & a_{M3} & \dots & a_{MN} \end{bmatrix}$$

3.3 Métricas de Qualidade Propostas

3.3.1 01) Rastreabilidade

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas:

$$R_n \implies \text{Rastreabilidade do atributo } n$$

Atribuição:

Atribuição:

$$R_n = \begin{cases} \text{Se } a_n \text{ é rastreável da captura até sua disponibilidade, } R_n = 1 \\ \text{Caso contrário, } R_n = 0 \end{cases}$$

Métrica:

$$RS = \frac{\sum_{n=1}^{N} R_n}{N}$$

Resultado: Se $RS = 1$, significa que o conjunto de dados é totalmente rastreável.

3.3.2 02) Integridade

Aplicação:

$$\forall a_{mn} \mid m = 1..M,\ n = 1..N$$

Variáveis específicas:

$$\begin{aligned} I_{mn} &\implies \text{Integridade do atributo } n \text{ na instância } m \\ I_n &\implies \text{Integridade total do atributo } n \end{aligned}$$

Atribuição:

$$I_{mn} = \begin{cases} \text{Se o hash de } a_{mn} \text{ é igual ao hash do dado original, } I_{mn} = 1 \\ \text{Caso contrário, } I_{mn} = 0 \end{cases}$$

Métrica:

$$I_n = \frac{\sum_{m=1}^{M} I_{mn}}{M}$$

$$IT = \frac{\sum_{n=1}^{N} I_n}{N}$$

Resultado: Se $IT = 1$, significa que o conjunto de dados possui totalmente preservada a sua integridade.

3.3.3 03) Privacidade

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas:

$$P_n \implies \text{Privacidade do atributo } n$$

Atribuição:

$$P_n = \begin{cases} \text{Se } a_n \text{ não é sensível à privacidade, } P_n = 1 \\ \text{Caso contrário, se } a_n \text{ permite uso com mascaramento, } P_n = 0.5 \\ \text{Caso contrário, } P_n = 0 \end{cases}$$

Métrica:

$$PV = \frac{\sum_{n=1}^{N} P_n}{N}$$

Resultado: Se $PV = 1$ significa que o conjunto de atributos não possui dados sensíveis à privacidade.

3.3.4 04) Disponibilidade

Aplicação:

$$\forall a_n \mid n = 1..N, \text{ considerando apenas atributos voláteis}$$

Variáveis específicas:

$$\begin{aligned} V &\implies \text{Quantidade de atributos voláteis em } A \\ D_n &\implies \text{Disponibilidade do atributo } n \text{ em local, formato e horários definidos} \end{aligned}$$

Atribuição:

$$D_n = \begin{cases} \text{Se } a_n \text{ é disponível, } D_n = 1 \\ \text{Caso contrário, } D_n = 0 \end{cases}$$

Métrica:

$$DS = \frac{\sum_{n=1}^{V} D_n}{V}$$

Resultado: Se $DS = 1$ significa que todos os atributos voláteis do conjunto de dados são totalmente disponíveis.

3.3.5 05) Relevância

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas

D_n	$\Longrightarrow$	Conjunto de valores possíveis para o atributo n
		$D_n = \{d_{n,1}, d_{n,2}, ..., d_{n,\|D_n\|}\}$
C	$\Longrightarrow$	Conjunto de rótulos de A
x	$\Longrightarrow$	Valor x, sendo $\{x \in D_n\}$
$P(x)$	$\Longrightarrow$	Probabilidade de ocorrer x

A Entropia do atributo a_n do conjunto A, com respeito ao atributo classe, representado por C, é dado por:

$$H_{(a_n;C)} = \sum_{i=1}^{|D_n|} P_{(x=d_{n,i})} H(d_{n,i}; C)$$
$$Onde:$$
$$H_{(d_{n,i};C)} = \sum_{j=1}^{|C|} P_{(x=d_{n,i};c_j)} I_{(x=d_{n,i};c_j)}$$
$$I_{(x=d_{n,i};c_j)} = \log_{10} \frac{1}{P_{(x=d_{n,i};c_j)}}$$

O conjunto ordenado de valores de entropias $\mathbb{H}$ calculado para cada atributo é definido como:

$$\mathbb{H} = \left\{ \; H_{(a_i;C)} \in \mathbb{R} | H_{(a_i;C)} < H_{(a_{i+1};C)} \; \right\}$$

Para o conjunto A, um atributo a_k, correspondente ao $minH_{(a_k;C)}$ pode representar um atributo fortemente relevante por estar mais relacionado diretamente com o atributo classe. O atributo correspondente ao $maxH_{(a_k;C)}$ pode representar um atributo fracamente relevante para a classificação. Os atributos com entropia entre os valores extremos podem corresponder a atributos relevantes. Por meio de inspeção humana, e com auxílio de especialista de domínio, é possível avaliar a relevância utilizando o conjunto $\mathbb{H}$:

Atribuição:

$$R_n = \begin{cases} \text{Se } a_n \text{ é fortemente relevante (menor entropia), } R_n = 1 \\ \text{Se } a_n \text{ é relevante, } R_n = 0.5 \\ \text{Se } a_n \text{ é fracamente relevante (maior entropia), } R_n = 0 \end{cases}$$

Métrica:

$$RL = \frac{\sum_{n=1}^{N} R_n}{N}$$

Resultado: Se $RL = 1$ significa que todos os atributos n são fortemente relevantes para os estudos a serem realizados no projeto. É importante ressaltar que um valor $RL = 1$ pode também indicar inúmeros atributos que favorecem a classificação direta, característica esta que deve ser comprovada por meio de uma análise de causalidade para evitar modelos "ingênuos", onde atributos fortemente relevantes podem corresponder a atributos de "efeito" e não "causais", situação esta que deve ser evitada.

3.3.6 06) Interpretabilidade

Aplicação:

$$\text{Aplicação: } \forall a_n \mid n = 1..N$$

Variáveis específicas

$$\begin{aligned} D_n &\implies \text{Domínio de valores de } n \\ I_n &\implies \text{Interpretabilidade do atributo } n \end{aligned}$$

Atribuição:

$$I_n = \begin{cases} \text{Se } D_n \text{ é interpretável, } I_n = 1 \\ \text{Caso contrário, } I_n = 0 \end{cases}$$

Métrica:

$$IP = \frac{\sum_{n=1}^{N} I_n}{N}$$

Resultado: Se $IP = 1$ significa que o conjunto de dados, na sua totalidade, é interpretável, ou seja, é possível compreender, em análise humana, a informação contida em todos os valores do atributo.

3.3.7 07) Consistência

3.3.7.1 Consistência Sintática

Aplicação:

$$\forall a_{mn} \mid m = 1..M, n = 1..N$$

Variáveis específicas:

$D_n \implies$ Domínio de valores do atributo a_n válidos sintaticamente
$C_n \implies$ Consistência Sintática do atributo n
$C_{mn} \implies$ Consistência Sintática do atributo n na instância m

Atribuição:

$$C_{mn} = \begin{cases} \text{Se } a_{mn} \in D_n, \text{ então } C_{mn} = 1 \\ \text{Caso contrário, } C_{mn} = 0 \end{cases}$$

Métrica:

$$C_n = \frac{\sum_{m=1}^{M} C_{mn}}{M}$$

$$CN = \frac{\sum_{n=1}^{N} C_n}{N}$$

Resultado: Se $CN = 1$ significa que o conjunto de dados, na sua totalidade, possui consistência sintática.

3.3.7.2 Consistência Semântica

Aplicação:

$$\forall a_m \mid m = 1..M$$

Variáveis específicas:

$R \implies$ Quantidade de regras de negócio aplicáveis a todas as instâncias m
$r \implies$ Número da regra de negócio, onde $r = 1..R$
$C_{mr} \implies$ Consistência Semântica da instância m para regra r
$C_r \implies$ Consistência Semântica de todas as instâncias para regra r

Atribuição:

$$C_{mr} = \begin{cases} \text{Se } a_m \text{ satistaz a regra } r, \text{ então } C_{mr} = 1 \\ \text{Caso contrário, } C_{mr} = 0 \end{cases}$$

Métrica:

$$C_r = \frac{\sum_{m=1}^{M} C_{mr}}{M}$$

$$CM = \frac{\sum_{r=1}^{R} C_r}{R}$$

Resultado: Se $CM = 1$ significa que o conjunto de dados, atendeu todas as regras de negócio R e, portanto, possui consistência semântica completa.

3.3.8 08) Diversidade

Aplicação:

$$\text{Aplicação: } \forall a_n \mid n = 1..N$$

Variáveis específicas

D_n	$\Longrightarrow$	Conjunto de valores distintos para o atributo n
		$D_n = \{d_{n,1}, d_{n,2}, ..., d_{n,\lvert D_n \rvert}\}$
$\lvert D_n \rvert$	$\Longrightarrow$	Quantidade de valores distintos do atributo n
T	$\Longrightarrow$	Quantidade de instâncias distintas possíveis para A

Atribuição:

$$T = \prod_{n=1}^{N} |D_n|$$

Métrica:

$$DV = \frac{M}{T}$$

Resultado: Se $DV = 1$ significa que o conjunto de dados possui todas as instâncias possíveis do espaço problema. Note que em bases de dados de alta dimensionalidade esse valor é praticamente inalcançável. Esta medida pode ser útil para avaliar o quanto o conjunto de dados representa e contém instâncias do domínio de problema. Note que esta métrica só leva em consideração a combinação de valores dos atributos e não uma análise amostral multivariada em relação à população em estudo.

3.3.9 09) Eficácia

Aplicação:

$$\forall a_m \mid m = 1..M$$

Variáveis específicas:

D_n	$\Longrightarrow$	Conjunto de valores distintos do atributo n
$\lvert D_n \rvert$	$\Longrightarrow$	Quantidade de valores distintos do atributo n
T	$\Longrightarrow$	Quantidade de instâncias distintas possíveis do conjunto de referência
x	$\Longrightarrow$	Instância distinta do conjunto de referência
P_x	$\Longrightarrow$	Probabilidade de ocorrência de x
I_x	$\Longrightarrow$	Informação no elemento x
H_T	$\Longrightarrow$	Entropia do conjunto de referência
H_A	$\Longrightarrow$	Entropia do conjunto A

Atribuição:

$$T = \prod_{n=1}^{N} |D_n|$$

A probabilidade de uma instância x do conjunto de referência ocorrer é dada por:

$$P_x = \frac{1}{T}$$

A quantidade de informação contida no evento x é dada por

$$I_x = \log_{10}(\tfrac{1}{P_x})$$
$$H_T = \sum_{x=1}^{T} P_x I_x \text{ como } P_1 = P_2 = ... = P_T = \tfrac{1}{T}$$
$$H_T = I_x = \log_{10} T$$

Para o conjunto A

$$H_A = \log_{10} M$$

Métrica:

$$EC = \log_{10} T - \log_{10} M = \log_{10} \frac{T}{M}$$

Resultado: Quanto menor for o valor de EC, significa que o conjunto de dados A contêm quantidade de informação próxima ao conjunto de referência (quando levado em consideração todos os valores possíveis de cada atributo). Note que, esta métrica pode ser aplicada durante um processo de seleção de atributos. Quando atributos são removidos do conjunto A, a tendência é aumentar a redundância de registros do conjunto. Isto diminuirá a entropia de A e a informação contida nele. Logo, isso pode afetar a eficácia na descoberta de novo conhecimento. Quanto mais próximo de zero EC estiver, maior será a eficácia do conjunto A.

3.3.10 10) Justiça

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas:

T	$\Longrightarrow$	Quantidade de atributos sensíveis a questões de justiça em A
Jn	$\Longrightarrow$	Índice de justiça do atributo a_n
$\|D_n\|$	$\Longrightarrow$	Quantidade de valores distintos do atributo a_n
O_i	$\Longrightarrow$	Frequência observada de cada valor do atributo a_n
E_i	$\Longrightarrow$	Frequência equiprovável esperada para cada valor do atributo a_n
P_i	$\Longrightarrow$	Probabilidade equiprovável esperada para cada valor do atributo a_n, sendo que $P_i = \frac{\|D_n\|}{M}$
γ	$\Longrightarrow$	Graus de liberdade
α	$\Longrightarrow$	Índice de significância de 5%

Atribuição:

O objetivo é verificar se cada atributo a_n, sensível às questões de justiça, possui uma distribuição uniforme. Pelo teste de aderência utilizando o método χ^2, tem-se o seguinte teste de hipóteses:

$\begin{cases} H_0\text{: distribuição da população é uniforme} \\ H_1\text{: tal não ocorre} \end{cases}$

Critério ou regra de decisão:
A hipótese H_0 será rejeitada se: $\chi^2_{calculado} > \chi^2_{critico}$

Onde: $\chi^2_{calculado} = \sum_{i=1}^{|D_n|} \frac{(O_i - E_i)^2}{E_i}$ com $E_i = M \times P_i$

$\chi^2_{critico} = \chi^2_{\gamma,\alpha}$ com $\gamma = |D_n| - 1$

Então: $J_n = \begin{cases} \text{Se a distribuição do atributo } a_n \text{ é uniforme, } J_n = 1 \\ \text{Caso contrário, } J_n = 0 \end{cases}$

Métrica:

$$JS = \frac{\sum_{j=1}^{T} J_n}{T}$$

Resultado: Se $JS = 1$ significa que todos os atributos analisados, sensíveis às questões de justiça, possuem distribuição uniforme, indicando que existem probabilidades semelhantes de ocorrência de valores em cada atributo dentro do conjunto de dados.

3.3.11 11) Representatividade

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas:

$R_n \implies$ Índice de Representatividade do atributo a_n
$D_n \implies$ Conjunto de valores distintos do atributo n
$|D_n| \implies$ Quantidade de valores distintos do atributo n
$O_i \implies$ Frequência observada de cada valor do atributo a_n
$E_i \implies$ Frequência esperada de cada valor do atributo a_n
$P_i \implies$ Probabilidade esperada para cada valor do atributo a_n
$\gamma \implies$ Graus de liberdade
$\alpha \implies$ Índice de significância de 5%

Atribuição:

O objetivo é verificar se cada atributo a_n, possui uma distribuição uniforme em relação à população.
Pelo teste de aderência utilizando o método χ^2, tem-se o seguinte teste de hipóteses:

$$\begin{cases} H_0\text{: distribuição da população é uniforme} \\ H_1\text{: tal não ocorre} \end{cases}$$

Critério ou regra de decisão:

$$\text{A hipótese } H_0 \text{ será rejeitada se: } \chi^2_{calculado} < \chi^2_{critico}$$

$$\text{Onde: } \chi^2_{calculado} = \sum_{i=1}^{|D_n|} \frac{(O_i - E_i)^2}{E_i} \text{ com } E_i = M \times P_i$$
$$\chi^2_{critico} = \chi^2_{\gamma,\alpha} \text{ com } \gamma = |D_n| - 1$$

$$\text{Então: } R_n = \begin{cases} \text{Se a distribuição do atributo } a_n \text{ é uniforme, } R_n = 1 \\ \text{Caso contrário, } R_n = 0 \end{cases}$$

Métrica:

$$RT = \frac{\sum_{j=1}^{N} R_j}{N}$$

Resultado: Se $RT = 1$ significa que todos os atributos analisados possuem distribuição proporcional ao universo observado.

3.3.12 12) Reprodutibilidade

Essa métrica é baseada no teste de χ^2 para a homogeneidade. O teste de homogeneidade verifica se um determinado atributo se distribui da mesma forma em várias amostras de uma ou várias populações. Construir uma tabela de contingências a partir da contagem de instâncias observadas (O) em C categorias de um atributo a_n (variável

aleatória categórica) para r amostras (ou populações) distintas.

Aplicação:

$$\forall a_n \mid n = 1..N$$

Variáveis específicas:

O_i	$\Longrightarrow$	Frequência observada de cada valor do atributo a_n
E_i	$\Longrightarrow$	Frequência esperada de cada valor do atributo a_n
Q	$\Longrightarrow$	Divergência entre as frequência esperada e observada
γ	$\Longrightarrow$	Graus de liberdade
α	$\Longrightarrow$	Índice de significância de 5%
R_n	$\Longrightarrow$	Reprodutibilidade do atributo n

Atribuição:

Amostra	1	2	3	...	c	*Total*
1	O_{11}	O_{12}	O_{13}	...	O_{1c}	n_1
2	O_{21}	O_{22}	O_{23}	...	O_{2c}	n_2
...	...	....	...	...	O_{1c}	n_1
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	n_r
Total	C_1	C_2	C_3	...	C_c	K

H_0: Todas as probabilidades para o mesmo valor categórico j são iguais.

$$p_1 j = p_2 j = ... = p_r j \forall j$$

H_a: Pelo menos uma das probabilidades p_{ij} é diferente de P_j para algum valor j. É testado se as frequências observadas diferem das frequências esperadas por meio do seguinte cálculo:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{|D_n|} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Onde: r, c = Número de amostras (populações) e valores categóricos do atributo n. O_{ij} = Frequência observada no atributo j da população i. E_{ij} = Frequência esperada no atributo j da população i.

$$E_{ij} = \frac{n_i C_j}{K}$$

Quanto maior o valor de Q maior será a divergência entre as frequências observadas e esperadas. Q possui distribuição aproximada χ^2 com $\gamma = (r-1) \times (|D_n| - 1)$

graus de liberdade.

Decisão do teste: se $O > \chi^2_\gamma; 1 - \alpha$ rejeita-se H_0 para o nível de significância α. Ao rejeitar H_0 se estabelece que o atributo a_n não se distribui da mesma forma em todas as amostras e, portanto, a reprodutibilidade pelo atributo a_n não é garantida.

$$R_n = \begin{cases} \text{Se } H_0 \text{ é aceita, } R_n = 1 \\ \text{Caso contrário, } R_n = 0 \end{cases}$$

Métrica:

$$RD = \frac{\sum_{n=1}^{N} R_n}{N}$$

Resultado: Se $RD = 1$ significa que todos os atributos possuem distribuição uniforme na amostra e, portanto, o *dataset A* é totalmente reprodutível.

3.4 Funções de Utilidade

Sendo o objetivo a construção de modelos de aprendizado adequados ao domínio de problema, alguns aspectos de qualidade podem ter maior ou menor utilidade do que outros aspectos. Para isso, é associado à métrica de qualidade de cada aspecto uma Função de Utilidade FU, que deve ser parametrizada de acordo com os objetivos que se deseje alcançar no projeto. Para isso, é proposto utilizar as funções Linear, Exponencial, Sigmoidal, e Sigmoidal Espelhada.

3.4.1 Definição das Variáveis

T	$\Longrightarrow$	Quantidade de aspectos avaliados no *dataset A*
X_j	$\Longrightarrow$	Valor da métrica de qualidade do aspecto j
FU_j	$\Longrightarrow$	Valor da função utilidade do aspecto j
W_j	$\Longrightarrow$	Peso do aspecto j sendo $W_j[0:1]$ e $\sum W_j = 1$
V_{min}	$\Longrightarrow$	Valor mínimo de qualidade aceito para o aspecto, que é definido conforme sua importância para o projeto
e	$\Longrightarrow$	Número de Euler, base da função exponencial natural
β	$\Longrightarrow$	Ajuste da relação exponencial da qualidade com a FU, sendo $5 <= \beta <= 20$

3.4.2 Função Utilidade Linear

A Função de Utilidade Linear é aplicada para os aspectos cuja evolução da utilidade pode ser considerada proporcional à evolução do valor da qualidade do aspecto X_j.

Considerando o Gráfico 1, o valor de V_{min} significa que para valores do índice de qualidade abaixo desse valor, o nível de qualidade do aspecto não é relevante para o projeto e, portanto, $FU_j = 0$. No Gráfico 1 pode ser visto que quanto maior V_{min} menos tolerante é o projeto à baixa qualidade do aspecto.

$$FU_j = \frac{1}{1 - V_{min}} X_j - \frac{V_{min}}{1 - V_{min}} = \frac{X_j - V_{min}}{1 - V_{min}}$$

3.4.3 Função Utilidade Exponencial

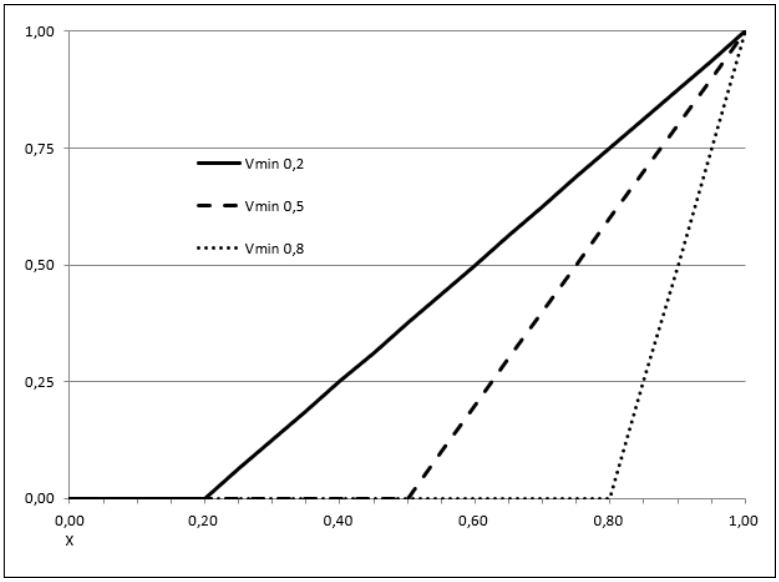
A Função Exponencial Natural é utilizada para os aspectos cujo aumento da utilidade possui relação com o aumento do valor da qualidade do aspecto X_j de forma variável. Inicialmente proporciona uma utilidade mais alta para baixos níveis de qualidade e, à medida em que o valor da qualidade vai aumentando, a utilidade vai alcançando seu máximo valor de 1. No Gráfico 2 pode ser visto esse comportamento.

$$FU_j = 1 - e^{-\beta X_j}$$

Dependendo do valor de β, diversas relações podem ser representadas. Por exemplo, para $\beta = 5$, um índice de qualidade de 0,30 já é suficiente para uma utilidade de 0,75 para o aspecto no projeto considerado.

Gráfico 1 – Gráfico Função Utilidade Linear

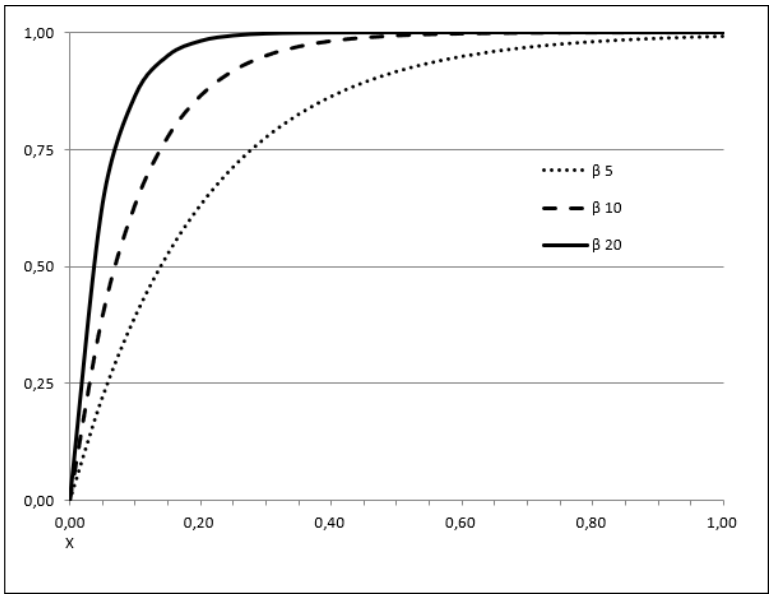
		Vmin		
		0,2	0,5	0,8
X	0,00	0,0000	0,0000	0,0000
	0,05	0,0000	0,0000	0,0000
	0,10	0,0000	0,0000	0,0000
	0,15	0,0000	0,0000	0,0000
	0,20	0,0000	0,0000	0,0000
	0,25	0,0625	0,0000	0,0000
	0,30	0,1250	0,0000	0,0000
	0,35	0,1875	0,0000	0,0000
	0,40	0,2500	0,0000	0,0000
	0,45	0,3125	0,0000	0,0000
	0,50	0,3750	0,0000	0,0000
	0,55	0,4375	0,1000	0,0000
	0,60	0,5000	0,2000	0,0000
	0,65	0,5625	0,3000	0,0000
	0,70	0,6250	0,4000	0,0000
	0,75	0,6875	0,5000	0,0000
	0,80	0,7500	0,6000	0,0000
	0,85	0,8125	0,7000	0,2500
	0,90	0,8750	0,8000	0,5000
	0,95	0,9375	0,9000	0,7500
	1,00	1,0000	1,0000	1,0000



Fonte: Elaborado pelo autor

Gráfico 2 – Gráfico Função Utilidade Exponencial

		β		
		5	10	20
X	0,00	0,0000	0,0000	0,0000
	0,05	0,2212	0,3935	0,6322
	0,10	0,3935	0,6322	0,8647
	0,15	0,5277	0,7769	0,9502
	0,20	0,6322	0,8647	0,9817
	0,25	0,7136	0,9180	0,9933
	0,30	0,7769	0,9502	0,9975
	0,35	0,8263	0,9698	0,9991
	0,40	0,8647	0,9817	0,9997
	0,45	0,8946	0,9889	0,9999
	0,50	0,9180	0,9933	1,0000
	0,55	0,9361	0,9959	1,0000
	0,60	0,9502	0,9975	1,0000
	0,65	0,9613	0,9985	1,0000
	0,70	0,9698	0,9991	1,0000
	0,75	0,9765	0,9994	1,0000
	0,80	0,9817	0,9997	1,0000
	0,85	0,9857	0,9998	1,0000
	0,90	0,9889	0,9999	1,0000
	0,95	0,9914	0,9999	1,0000
	1,00	0,9933	1,0000	1,0000



Fonte: Elaborado pelo autor

3.4.4 Função Utilidade Sigmoidal

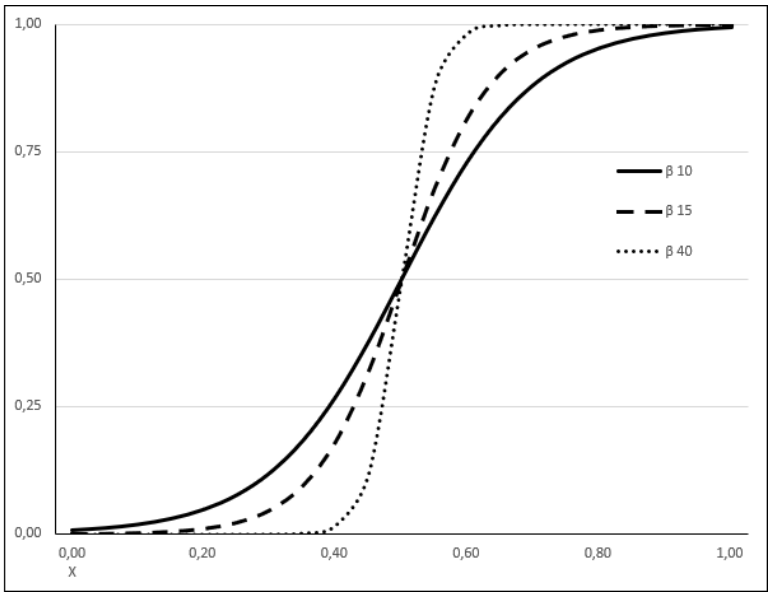
A Função Sigmoidal é utilizada para os aspectos cujo aumento da utilidade possui relação com o aumento do valor da qualidade X_j também de forma variável. Inicialmente proporciona um valor de utilidade e taxa de variação baixas para baixos níveis de qualidade e, à medida em que a qualidade vai aumentando, chegando a valores próximos de $X_j = 0{,}5$, a taxa de variação da utilidade aumenta significativamente, voltando a reduzir essa taxa quando a qualidade atinge valores mais altos, próximos de 1. No Gráfico 3 pode ser visto esse comportamento.

$$FU_j = \frac{1}{1 + e^{-\beta(X_j - 0{,}5)}}$$

Dependendo do valor de β, diversas relações podem ser representadas. Por exemplo, para $\beta = 10$, um índice de qualidade de 0,10 apresenta uma utilidade de apenas 0,018, mas um índice 0,50 já apresenta utilidade de 0,5, para o aspecto no projeto considerado.

Gráfico 3 – Gráfico Função Utilidade Sigmoidal

		β		
		10	15	40
X	0,00	0,0067	0,0006	0,0000
	0,05	0,0110	0,0012	0,0000
	0,10	0,0180	0,0025	0,0000
	0,15	0,0293	0,0052	0,0000
	0,20	0,0474	0,0110	0,0000
	0,25	0,0759	0,0230	0,0000
	0,30	0,1192	0,0474	0,0003
	0,35	0,1824	0,0953	0,0025
	0,40	0,2689	0,1824	0,0180
	0,45	0,3775	0,3208	0,1192
	0,50	0,5000	0,5000	0,5000
	0,55	0,6225	0,6792	0,8808
	0,60	0,7311	0,8176	0,9820
	0,65	0,8176	0,9047	0,9975
	0,70	0,8808	0,9526	0,9997
	0,75	0,9241	0,9770	1,0000
	0,80	0,9526	0,9890	1,0000
	0,85	0,9707	0,9948	1,0000
	0,90	0,9820	0,9975	1,0000
	0,95	0,9890	0,9988	1,0000
	1,00	0,9933	0,9994	1,0000



Fonte: Elaborado pelo autor

3.4.5 Função Utilidade Sigmoidal Espelhada

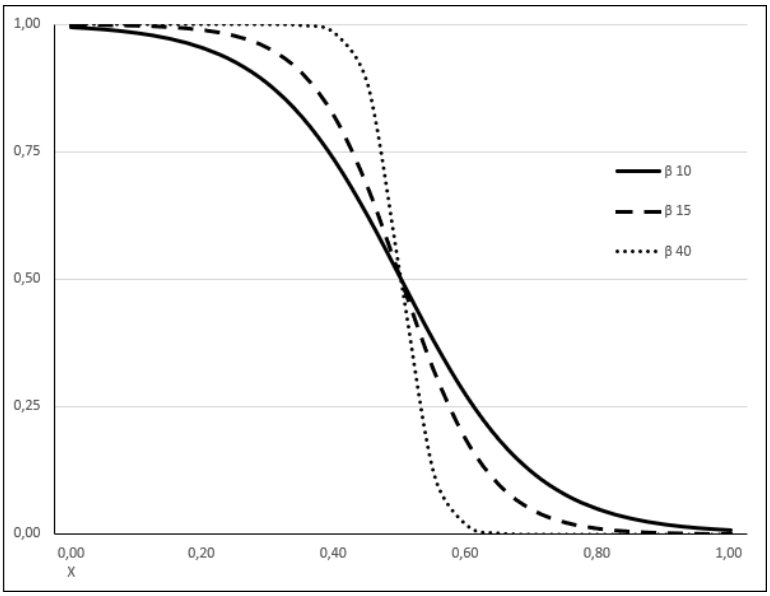
Baseada na Função Sigmoidal, pode ser utilizada para aspectos cujas medidas apresentam valores invertidos para expressar a qualidade (quando $X_j = 0$: melhor qualidade, e para $X_j > 0$: a qualidade gradativamente vai diminuindo). No Gráfico 4 pode ser visto este comportamento.

$$FU_j = 1 - \frac{1}{1 + e^{\beta(X_j - 0{,}5)}}$$

Dependendo do valor de β, diversas relações podem ser representadas. Por exemplo, para $\beta = 10$, um índice de qualidade de 0,90 apresenta uma utilidade de apenas 0,018, mas um índice 0,10 já apresenta utilidade de 0,982 para o aspecto no projeto considerado.

Gráfico 4 – Gráfico Função Utilidade Sigmoidal Espelhada

		β		
		10	15	40
X	0,00	0,9933	0,9994	1,0000
	0,05	0,9890	0,9988	1,0000
	0,10	0,9820	0,9975	1,0000
	0,15	0,9707	0,9948	1,0000
	0,20	0,9526	0,9890	1,0000
	0,25	0,9241	0,9770	1,0000
	0,30	0,8808	0,9526	0,9997
	0,35	0,8176	0,9047	0,9975
	0,40	0,7311	0,8176	0,9820
	0,45	0,6225	0,6792	0,8808
	0,50	0,5000	0,5000	0,5000
	0,55	0,3775	0,3208	0,1192
	0,60	0,2689	0,1824	0,0180
	0,65	0,1824	0,0953	0,0025
	0,70	0,1192	0,0474	0,0003
	0,75	0,0759	0,0230	0,0000
	0,80	0,0474	0,0110	0,0000
	0,85	0,0293	0,0052	0,0000
	0,90	0,0180	0,0025	0,0000
	0,95	0,0110	0,0012	0,0000
	1,00	0,0067	0,0006	0,0000



Fonte: Elaborado pelo autor

3.4.6 Função Utilidade para cada aspecto de qualidade

Na Tabela 3 pode ser vista a recomendação da Função Utilidade que pode ser aplicada a cada aspecto de qualidade.

3.4.7 Função Utilidade Total

A Função Utilidade Total, que pondera a utilidade de todos os aspectos avaliados, fornecendo uma qualidade total do conjunto de dados, pode ser sintetizada como:

$$FU_t = \sum_{j=1}^{T} W_j FU_j$$

Para cada aspecto j sendo $W_j[0:1]$ e $\sum W_j = 1$

Tabela 3 – Aplicação da Função Utilidade aos aspectos

Aspecto	Li	Ex	Si	Se	Justificativa
1) Rastreabilidade		X			Rastreabilidade baixa não impede a construção do modelo de aprendizado. Portanto, definir um valor de FU relativamente alto seria aceito. e.g. Para $\beta = 10$, uma rastreabilidade de 0,30 proporciona um $FU = 0{,}95$.
2) Integridade	X				Domínios de alta criticidade podem demandar exatidão no valor do atributo. Qualquer alteração no valor pode prejudicar a representatividade do modelo. e.g. Para uma integridade de 0,85 e $V_{min} = 0{,}8$, $FU = 0{,}25$.
3) Privacidade	X				Privacidade de valor baixo pode comprometer a legalidade e aplicação do modelo. e.g. Para uma privacidade garantida de 0,95 e $V_{min} = 0{,}8$, $FU = 0{,}75$.
4) Disponibilidade		X			Disponibilidade baixa não impede a construção do modelo de aprendizado. Portanto, definir um valor de FU relativamente alto seria aceito. e.g. Para $\beta = 20$, uma disponibilidade de 0,10 proporciona um $FU = 0{,}86$.
5) Relevância	X				É necessário um alto índice de qualidade desse aspecto para evitar a obtenção de modelos fracos para o aprendizado desejado e pouco representativos do domínio e.g. Para $V_{min} = 0{,}5$, $FU = 0{,}60$ se a relevância for 0,80.
6) Interpretabilidade		X			Interpretabilidade baixa não compromete significativamente a construção do modelo. Há maior tolerância com índices baixos e, portanto utilizada a FU exponencial. e.g. Para $\beta = 20$, uma disponibilidade de 0,10 proporciona um $FU = 0{,}86$.
7) Consistência	X				Tanto na consistência sintática quanto na semântica um índice baixo de qualidade provavelmente significa menor confiabilidade da amostra, comprometendo o modelo a ser construído e.g. Para $V_{min} = 0{,}8$, FU será 0,75 se a consistência for 0,95.
8) Diversidade			X		Para obter modelos representativos do domínio de problema, o número de instâncias únicas deve ser suficientemente grande para cobrir a maior parte do espaço de possibilidades. Diversidades com valor muito baixo (típico em projetos de aprendizado de máquina), devem apresentar utilidades também baixas. e.g. Para $\beta = 10$, e um índice de diversidade de 0,1, $FU = 0{,}018$.
9) Eficácia				X	A Eficácia compara a quantidade de informação contida entre a amostra e o espaço de possibilidades. Uma eficácia próxima de zero indica que a amostra contêm a mesma quantidade de informações que a do espaço de possibilidade. Nesse caso $FU = 1$. À medida em que o valor da eficácia aumenta, indica que a amostra possui baixa utilidade. e.g. Para $\beta = 10$, quando $T = 10 * M$, a eficácia = 1, e $FU = 0{,}0$.
10) Justiça	X				Uma vez que existe atualmente uma grande preocupação em evitar que modelos de aprendizado sejam tendenciosos e, por consequência, discriminatórios, a utilidade desse aspecto considera crítica a sua qualidade, usando a FU linear e não aceitando V_{min} menor que 0,5. e.g. para um índice de justiça de 0,75, considerando $V_{min} = 0{,}5$, $FU = 0{,}50$.
11) Representatividade			X		A utilidade do modelo tem grande dependência da qualidade desse aspecto porque ele está fortemente ligado ao quanto o estudo feito com a amostra é aplicável ao universo estudado. Portanto, não é permissível ser tolerante com qualidade fraca para esse aspecto (típico em projetos de Aprendizado de Máquina). e.g. Para $\beta = 10$, e um índice de representatividade de 0,40, ainda terá uma utilidade relativamente baixa $FU = 0{,}27$.
12) Reprodutibilidade		X			Reprodutibilidade baixa não impede a construção do modelo de aprendizado, porém restrições do modelo devem ser impostas, desde que a amostra considerada impôs restrições. e.g. Para $\beta = 10$, uma reprodutibilidade de 0,25 proporcionaria uma $FU = 0{,}91$.

Li $\Longrightarrow$ Função Utilidade Linear
Ex $\Longrightarrow$ Função Utilidade Exponencial
Si $\Longrightarrow$ Função Utilidade Sigmoidal
Se $\Longrightarrow$ Função Utilidade Sigmoidal Espelhada

Fonte: Elaborada pelo autor

4 APLICAÇÃO DA PROPOSTA DE MÉTRICAS - CONJUNTO DE DADOS TOY

Nesta seção será feita a aplicação das métricas propostas, para demonstração do seu funcionamento e de como obter o índice de qualidade do *dataset* do projeto. O *dataset* utilizado é uma pequena amostra fictícia de dados (toy), apenas para fins de exemplo, e pode ser visto na Tabela 4, assim como sua discretização e definição de domínios de atributos podem ser vistas na Tabela 5.

Tabela 4 – Conjunto de dados toy - Apresentação

Nome	**Sexo a_1**	**Idade a_2**	**Ocupação a_3**	**Escolaridade a_4**	**Salário a_5**	**Filhos a_6**	**Padrão Consumo *classe***
João m_1	M	39 Adulto	Analista	Superior	9000 Alta	2 Dois	Alto
José m_2	M	81 Idoso	Estagiário	Médio	750 Baixa	0 Nenhum	Baixo
Laura m_3	F	16 Adolesc	Estagiário	Superior	600 Baixa	0 Nenhum	Baixo
Paulo m_4	M	19 Adolesc	Estagiário	Médio	700 Baixa	1 Um	Alto
Lucas m_5	M	22 Adulto	Estagiário	Superior	6100 Alta	1 Um	Baixo
Eduardo m_6	M	18 Adolesc	Analista	Médio	1600 Baixa	0 Nenhum	Baixo
Felipe m_7	M	18 Adolesc	Estagiário	Fundamental	2600 Média	2 Dois	Baixo
Pedro m_8	M	23 Adulto	Gerente	Médio	2500 Média	0 Nenhum	Baixo
Maria m_9	F	29 Adulto	Gerente	Superior	15000 Alta	2 Dois	Alto
Bruna m_{10}	F	71 Idoso	Gerente	Fundamental	3500 Média	0 Nenhum	Alto

Fonte: Elaborado pelo autor

Tabela 5 – Conjunto de dados toy - Domínios e Discretizações de Valores

$D_{n1} \implies$ 2, sendo: M = Masculino e F = Feminino
$D_{n2} \implies$ 3, sendo: A = Adolescente (12-20), U = Adulto (21-64) e S = Idoso (65+)
$D_{n3} \implies$ 3, sendo: E = Estagiário, A = Analista e G = Gerente
$D_{n4} \implies$ 3, sendo: F = Fundamental, M = Médio e S = Superior
$D_{n5} \implies$ 3, sendo: B = Baixa (0-2000), M = Média (2001-5000) e A = Alta (5001+)
$D_{n6} \implies$ 4, sendo: N = Nenhum (0), U = Um (1), D = Dois (2) e M = Mais (3+)

Fonte: Elaborado pelo autor

4.1 Cálculo dos Índices de Qualidade

1) *Rastreabilidade:* será considerado que somente o atributo Salário (a_5) não é Rastreável, ficando então $R_5 = 0$. Para todos os demais atributos $R_n = 1$. Então:

$$RS = \frac{1+1+1+1+0+1}{6} = \frac{5}{6} = 0{,}833$$

2) *Integridade:* assume-se que não há possibilidade de erro no processo de extração/-transformação até compor o conjunto de dados. Para cada atributo a_n a integridade será

$I_n = 1 \ \forall a_n \mid n = 1..6$. Isto ocorre por que os atributos do conjunto toy não são de alta precisão e não requerem truncamento, ou arrendondamento dos dados. Portanto, a métrica de integridade do conjunto de dados será $ITS = 1$.

3) *Privacidade:* considerando que apenas a_5 é sensível à privacidade, pois não possui restrição de acesso, mas exige o uso com mascaramento, $P_{a_5} = 0{,}5$. Para o demais atributos, $P = 1$. Então: $PV = (1 + 1 + 1 + 1 + 0{,}5 + 1)/6 = 0{,}917$

4) *Disponibilidade:* dois atributos voláteis podem influenciar o padrão de consumo obtido a partir do modelo de aprendizado: Salário (a_5) e Quantidade de filhos (a_6), sendo então $V = 2$. Considerando a disponibilidade para extração desses atributos $D_5 = D_6 = 1$. Portanto, $DS = 1$.

5) Relevância Para cada atributo a_n é calculada a entropia referente ao atributo classe:

$$
\begin{aligned}
H_{(a_1;C)} &= \tfrac{7}{10} * H_{(M;C)} + \tfrac{3}{10} * H_{(F;C)} \\
&= \tfrac{7}{10} * \left[\ \tfrac{2}{7} * \log_{10}\left(\tfrac{7}{2}\right) + \tfrac{5}{7} * \log_{10}\left(\tfrac{7}{5}\right) \ \right] + \tfrac{3}{10} * \left[\ \tfrac{2}{3} * \log_{10}\left(\tfrac{3}{2}\right) + \tfrac{1}{3} * \log_{10}\left(\tfrac{3}{1}\right) \ \right] \\
&= \tfrac{7}{10} * \left[\ \tfrac{2}{7} * 0{,}544 + \tfrac{5}{7} * 0{,}146 \ \right] + \tfrac{3}{10} * \left[\ \tfrac{2}{3} * 0{,}176 + \tfrac{1}{3} * 0{,}477 \ \right] \\
&= \tfrac{7}{10} * \left[\ \tfrac{1{,}818}{7} + \tfrac{0{,}73}{7} \ \right] + \tfrac{3}{10} * \left[\ \tfrac{0{,}352}{3} + \tfrac{0{,}477}{3} \ \right] \\
&= \tfrac{7}{10} * \tfrac{1{,}818}{7} + \tfrac{3}{10} * \tfrac{0{,}829}{3} \\
&= 0{,}265
\end{aligned}
$$

$$
\begin{aligned}
H_{(a_2;C)} &= \tfrac{4}{10} * H_{(A;C)} + \tfrac{4}{10} * H_{(U;C)} + \tfrac{2}{10} * H_{(S;C)} \\
&= \tfrac{4}{10} * \left[\ \tfrac{1}{4} * \log_{10}\left(\tfrac{4}{1}\right) + \tfrac{3}{4} * \log_{10}\left(\tfrac{4}{3}\right) \ \right] + \tfrac{4}{10} * \left[\ \tfrac{2}{4} * \log_{10}\left(\tfrac{4}{2}\right) + \tfrac{2}{4} * \log_{10}\left(\tfrac{4}{2}\right) \ \right] + \\
&\quad \tfrac{2}{10} * \left[\ \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) + \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) \ \right] \\
&= 0{,}278
\end{aligned}
$$

$$
\begin{aligned}
H_{(a_3;C)} &= \tfrac{5}{10} * H_{(E;C)} + \tfrac{2}{10} * H_{(A;C)} + \tfrac{3}{10} * H_{(G;C)} \\
&= \tfrac{5}{10} * \left[\ \tfrac{1}{5} * \log_{10}\left(\tfrac{5}{1}\right) + \tfrac{4}{5} * \log_{10}\left(\tfrac{5}{4}\right) \ \right] + \tfrac{2}{10} * \left[\ \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) + \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) \ \right] + \\
&\quad \tfrac{3}{10} * \left[\ \tfrac{2}{3} * \log_{10}\left(\tfrac{3}{2}\right) + \tfrac{1}{3} * \log_{10}\left(\tfrac{3}{1}\right) \ \right] \\
&= 0{,}252
\end{aligned}
$$

$$
\begin{aligned}
H_{(a_4;C)} &= \tfrac{2}{10} * H_{(F;C)} + \tfrac{4}{10} * H_{(M;C)} + \tfrac{4}{10} * H_{(S;C)} \\
&= \tfrac{2}{10} * \left[\ \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) + \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) \ \right] + \tfrac{4}{10} * \left[\ \tfrac{1}{4} * \log_{10}\left(\tfrac{4}{1}\right) + \tfrac{3}{4} * \log_{10}\left(\tfrac{4}{3}\right) \ \right] + \\
&\quad \tfrac{4}{10} * \left[\ \tfrac{2}{4} * \log_{10}\left(\tfrac{4}{2}\right) + \tfrac{2}{4} * \log_{10}\left(\tfrac{4}{2}\right) \ \right] \\
&= 0{,}278
\end{aligned}
$$

$$
\begin{aligned}
H_{(a_5;C)} &= \tfrac{4}{10} * H_{(B;C)} + \tfrac{3}{10} * H_{(M;C)} + \tfrac{3}{10} * H_{(A;C)} \\
&= \tfrac{4}{10} * \left[\ \tfrac{1}{4} * \log_{10}\left(\tfrac{4}{1}\right) + \tfrac{3}{4} * \log_{10}\left(\tfrac{4}{3}\right) \ \right] + \tfrac{3}{10} * \left[\ \tfrac{1}{3} * \log_{10}\left(\tfrac{3}{1}\right) + \tfrac{2}{3} * \log_{10}\left(\tfrac{3}{2}\right) \ \right] + \\
&\quad \tfrac{3}{10} * \left[\ \tfrac{2}{3} * \log_{10}\left(\tfrac{3}{2}\right) + \tfrac{1}{3} * \log_{10}\left(\tfrac{3}{1}\right) \ \right] \\
&= 0{,}264 \\
H_{(a_6;C)} &= \tfrac{5}{10} * H_{(T;C)} + \tfrac{2}{10} * H_{(U;C)} + \tfrac{3}{10} * H_{(V;C)} \\
&= \tfrac{5}{10} * \left[\ \tfrac{1}{5} * \log_{10}\left(\tfrac{5}{1}\right) + \tfrac{4}{5} * \log_{10}\left(\tfrac{5}{4}\right) \ \right] + \tfrac{2}{10} * \left[\ \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) + \tfrac{1}{2} * \log_{10}\left(\tfrac{2}{1}\right) \ \right] + \\
&\quad \tfrac{3}{10} * \left[\ \tfrac{2}{3} * \log_{10}\left(\tfrac{3}{2}\right) + \tfrac{1}{3} * \log_{10}\left(\tfrac{3}{1}\right) \ \right] + \tfrac{0}{10} * \left[\ \tfrac{0}{0} * \log_{10}\left(\tfrac{0}{0}\right) + \tfrac{0}{0} * \log_{10}\left(\tfrac{0}{0}\right) \ \right] \\
&= 0{,}252 \\
\mathbb{H} &= \left\{ \ a_3, a_6, a_5, a_1, a_2, a_4 \ \right\} = \left\{ \ 0.252, 0.252, 0.264, 0.265, 0.2778, 0.278 \ \right\}
\end{aligned}
$$

Como as entropias de todos os atributos a_1 a a_6 ficaram próximas, serão considerados fortemente relevantes e, portanto, para todos eles $R = 1$.

Então:

$$RL = \frac{R_{a1}+R_{a2}+R_{a3}+R_{a4}+R_{a5}+R_{a6}}{N} = \frac{1+1+1+1+1+1}{6} = \frac{6}{6} = 1$$

6) ***Interpretabilidade:*** para cada atributo a_n existe conceitualmente um conjunto domínio de valores possíveis D_n. Por exemplo para $D_4 = \{$Fundamental, Médio, Superior$\}$. Pela compreensão humana todos os valores do atributo possuem informação interpretável. Estendendo para todo o conjunto, a interpretabilidade será $I_n = 1 \ \forall a_n \mid n = 1..6$. Portanto, a interpretabilidade do conjunto de dados será $IP = 1$.

7.1) ***Consistência - Sintática:*** considerando que o domínio do atributo *idade*, (a_2) é $D_2 = [18,...,60]$, e que a_2 seja o único atributo que possui valor fora do seu domínio. Então, $C_n = 1$, exceto para o atributo a_2. A consistência sintática do atributo e do conjunto toy é dada por: $C_2 = (1 + 0 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1)/10 = 0{,}9$ e por $CN = (1 + 0{,}9 + 1 + 1 + 1 + 1)/6 = 0{,}98$

7.2) ***Consistência - Semântica:*** considerando duas regras semânticas $(R = 2)$: Regra 1: *Menores de 18 anos não podem ter nível de escolaridade Superior* e Regra 2: *Estagiários não podem ganhar mais que 1500.* Observar que m_3 e m_7 não atendem as regras respectivamente. Sendo assim, a consistência semântica devido a cada regra e do conjunto completo, é dado por: $C_{Regra1} = (1 + 1 + 0 + 1 + 1 + 1 + 1 + 1 + 1 + 1)/10 = 0{,}9$, $C_{Regra2} = (1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 1 + 1)/10 = 0{,}9$. $CM = (0{,}9 + 0{,}9)/2 = 0{,}9$

8) ***Diversidade:*** considerando a quantidade de valores possíveis dentro do domínio de cada atributo a_n, representado por $|D_n|$, $\forall a_n \mid n = 1..6$, ver Tabela 5, o total de instâncias

que descrevem completamente o domínio é dado por: $T = 2 \times 3 \times 3 \times 3 \times 3 \times 4 = 864$. A diversidade do conjunto de dados é dado por $DV = \frac{10}{864} = 0{,}015$ (1,5%). Esta métrica proporciona uma ideia da porcentagem de instâncias frente ao total de possibilidades do universo de estudo.

9) *Eficácia:* a eficácia do conjunto de dados A para representar todo o universo de possibilidade do domínio é medido por meio da variação $EC = \log_{10} 864 - \log_{10} 6 = 2.1584$. Como nesse caso EC ficou muito distante de zero, inclusive ultrapassando 1, é considerando que $EC = 0$.

10) *Justiça:* apenas os atributos a_1, a_2, a_4 e a_5 são sensíveis às questões de Justiça. Para o tributo a_1:

$$\chi^2_{calculado}(a_1) = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} = 1{,}6$$

$$\chi^2_{critico}(a_1) = \chi^2_{\gamma;\alpha} \text{ com } \gamma = 2 - 1 = 1 \implies \chi^2_{1;0,05} = 3{,}841$$

Como 1,6 não é maior que 3,841, H_0 não é rejeitada e considera-se que a distribuição de a_1 é uniforme. Portanto, $J_{a_1} = 1$.

Nos atributos a_2, a_4 e a_5 ocorre a proporção similar à do atributo a_1 e, com isso, todos terão $J = 1$. Sendo assim:

$$JS = \frac{J_{a_1} + J_{a_2} + J_{a_4} + J_{a_5}}{J} = \frac{1+1+1+1}{4} = \frac{4}{4} = 1$$

11) *Representatividade:* para essa aplicação experimental será considerado que o universo U estudado tem uma quantidade similar de ocorrências para cada valor do domínio. Então, para o atributo a_2:

$$\chi^2_{calculado}(a_2) = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(4-\frac{10}{3})^2}{\frac{10}{3}} + \frac{(4-\frac{10}{3})^2}{\frac{10}{3}} + \frac{(2-\frac{10}{3})^2}{\frac{10}{3}} = 0{,}79$$

$$\chi^2_{critico}(a_2) = \chi^2_{\gamma;\alpha} \text{ com } \gamma = 3 - 1 = 2 \implies \chi^2_{2;0,05} = 5{,}991$$

Como 0,79 não é maior que 5,991, H_0 não é rejeitada e considera-se que a distribuição de a_2 é uniforme. Portanto, $R_{a_2} = 1$.

Em todos os demais atributos (a_1, a_3, a_4, a_5 e a_6) ocorre a proporção similar à do atributo a_2 e, com isso, todos terão $R = 1$. Sendo assim:

$$RT = \frac{1+1+1+1+1+1}{6} = \frac{6}{6} = 1$$

12) ***Reprodutibilidade:*** Tratando-se de um exemplo toy, todas as amostras são equivalentes e portanto garantem a reprodutibilidade do modelo. Logo, $R_n = 1$ para todos os atributos. Então:

$$RD = \frac{\sum_{n=1}^{N} R_n}{N} = \frac{1+1+1+1+1+1}{6} = 1$$

4.2 Cálculo da Função Utilidade

1) ***Rastreabilidade:*** será utilizada a função exponencial e considerado $\beta = 10$:

$$FU_1 = 1 - 2{,}71882818^{-(10\times 0{,}833)} = 1 - \left(\frac{1}{2{,}71882818}\right)^{8{,}33} = 1 - 0{,}0002 = 0{,}9998$$

2) ***Integridade:*** será utilizada a função linear e considerado $V_{min} = 0{,}5$:

$$FU_2 = \frac{1 - 0{,}5}{1 - 0{,}5} = \frac{0{,}5}{0{,}5} = 1$$

3) ***Privacidade:*** será utilizada a função linear e considerado $V_{min} = 0{,}5$:

$$FU_3 = \frac{0{,}917 - 0{,}5}{1 - 0{,}5} = \frac{0{,}417}{0{,}5} = 0{,}834$$

4) ***Disponibilidade:*** será utilizada a função exponencial e considerado $\beta = 10$:

$$FU_4 = 1 - 2{,}71882818^{-(10\times 1)} = 1 - \left(\frac{1}{2{,}71882818}\right)^{10} = 1 - 0 = 1$$

5) ***Relevância:*** será utilizada a função linear e considerado $V_{min} = 0{,}5$:

$$FU_5 = \frac{1 - 0{,}5}{1 - 0{,}5} = \frac{0{,}5}{0{,}5} = 1$$

6) ***Interpretabilidade:*** será utilizada a função exponencial e considerado $\beta = 10$:

$$FU_6 = 1 - 2{,}71882818^{-(10\times 1)} = 1 - \left(\frac{1}{2{,}71882818}\right)^{10} = 1 - 0 = 1$$

7) ***Consistência - Sintática:*** será utilizada a função linear e considerado $V_{min} = 0{,}5$:

$$FU_{7.1} = \frac{0{,}98 - 0{,}5}{1 - 0{,}5} = \frac{0{,}48}{0{,}5} = 0{,}96$$

Consistência - Semântica: será utilizada a função exponencial e considerado $\beta = 10$:

$$FU_{7.2} = 1 - 2{,}71882818^{-(10\times0{,}9)} = 1 - \left(\frac{1}{2{,}71882818}\right)^{9} = 1 - 0{,}0001 = 0{,}9999$$

8) ***Diversidade:*** será utilizada a função sigmoidal e considerado $\beta = 15$:

$$FU_{8} = \frac{1}{1 + 2{,}71882818^{-15(0{,}015-0{,}5)}} = \frac{1}{1 + 2{,}71882818^{7{,}275}} = \frac{1}{1444{,}75119} = 0{,}0007$$

9) ***Eficácia:*** será utilizada a função sigmoidal espelhada e considerado $\beta = 15$:

$$FU_{10} = 1 - \frac{1}{1 + 2{,}71882818^{-15(0-0{,}5)}} = 1 - \frac{1}{1 + 2{,}71882818^{7{,}5}} = 1 - \frac{1}{1809{,}04241} = 0{,}9994$$

10) ***Justiça:*** será utilizada a função linear e considerado $V_{min} = 0{,}5$:

$$FU_{10} = \frac{1 - 0{,}5}{1 - 0{,}5} = \frac{0{,}5}{0{,}5} = 1$$

11) ***Representatividade:*** será utilizada a função sigmoidal e considerado $\beta = 15$:

$$FU_{11} = \frac{1}{1 + 2{,}71882818^{-15(0{,}8-0{,}5)}} = \frac{1}{1 + 2{,}71882818^{-4{,}5}} = \frac{1}{1{,}01111} = 0{,}989$$

12) ***Reprodutibilidade:*** será utilizada a função exponencial e considerado $\beta = 10$:

$$FU_{12} = 1 - 2{,}71882818^{-(10\times1)} = 1 - \left(\frac{1}{2{,}71882818}\right)^{10} = 1 - 0 = 1$$

Utilidade Total: na ponderação dos aspectos será considerado o mesmo peso para todos

os aspectos avaliados ($\frac{1}{13}$). Então:

$$\begin{aligned} FU_t &= 0{,}05 \times 0{,}9998 + 0{,}05 \times 1{,}0000 + 0{,}05 \times 0{,}8340 + 0{,}05 \times 1{,}0000 + 0{,}10 \times 1{,}0000 + \\ &\quad 0{,}05 \times 1{,}0000 + 0{,}10 \times 0{,}9600 + 0{,}10 \times 0{,}9999 + 0{,}10 \times 0{,}0007 + 0{,}10 \times 0{,}9994 + \\ &\quad 0{,}10 \times 1{,}0000 + 0{,}10 \times 0{,}9890 + 0{,}05 \times 1{,}0000 = 0{,}8866 \end{aligned}$$

5 APLICAÇÃO DA PROPOSTA DE MÉTRICAS - *DATASETS* DO UCI

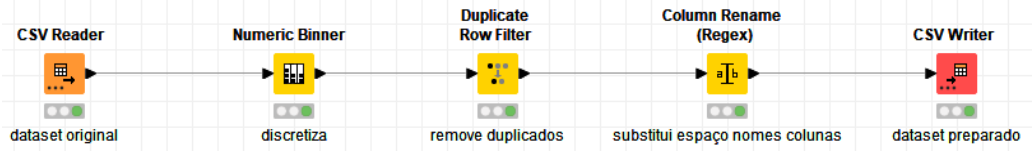
Neste capítulo as medidas de qualidade serão aplicados em bases de dados reais disponíveis no repositório University of California, Irvine (UCI) Machine Learning Repository (UCI, 2025). Atributos contínuos foram discretizados considerando intervalos com tamanho pré-definidos por inspeção humana. As bases de dados, tanto as originais quanto as resultantes após seus respectivos pré-processamentos, encontram-se no https://github.com/ericksonrsa/ucids. Tanto para os pré-processamentos dos *datasets* quanto para suas utilizações em um algoritmo de árvore de decisão foi utilizado o Knime, que é uma plataforma para Ciência de Dados de código aberto.

5.1 Seleção e Pré-processamento dos *Datasets*

A seguir está a apresentação de cada Dataset (DS) selecionado para a aplicação da proposta de métricas e os pré-processamentos realizados:

DS1) Iris: contém 150 instâncias com 4 atributos contínuos e 3 classes (Iris Setosa, Iris Versicolour e Iris Virginica). A base de dados contém 4 medidas características das plantas que são utilizadas para definir sua classe. Esta base de dados foi submetida a um processo de discretização de 4 intervalos para cada atributo numérico, sendo as faixas de valores de cada intervalo definidas de forma a ter quantidades de registros similares em cada intervalo. Após o processo de discretização as instâncias repetidas foram removidas do conjunto de dados para cálculo das medidas, restando 54 registros. Esse processo de discretização e posterior remoção de registros duplicados foi realizado com a utilização do Knime e o fluxo pode ser visto na Figura 1.

Figura 1 – Discretização do *dataset* Iris

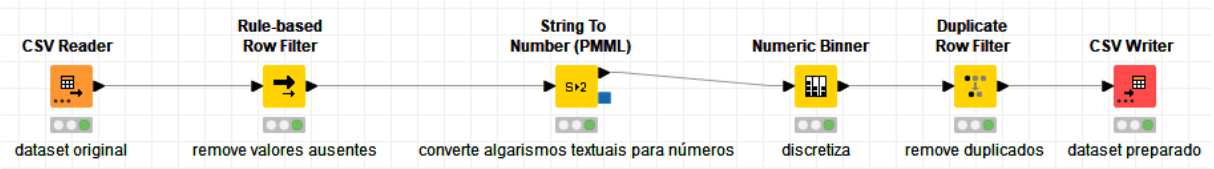


Fonte: Elaborado pelo autor

DS2) Doenças Cardíacas: contém 303 instâncias com 13 atributos (6 contínuos e 7 categóricos) e 2 classes(Não possui e Possui doença cardíaca). Um dos mais populares e amplamente utilizados para tarefas de classificação. Ele contém informações sobre pacientes e uma série de atributos médicos que podem ser usados para prever a presença

de doenças cardíacas. Foram removidas as 6 instâncias que continham valores ausentes, restando 297, que tiveram seus atributos contínuos discretizados. Em seguida foi realizada a remoção das instâncias repetidas, restando então 279 registros. Esse processo de discretização e posterior remoção de registros duplicados foi realizado com a utilização do Knime e o fluxo pode ser visto na Figura 2.

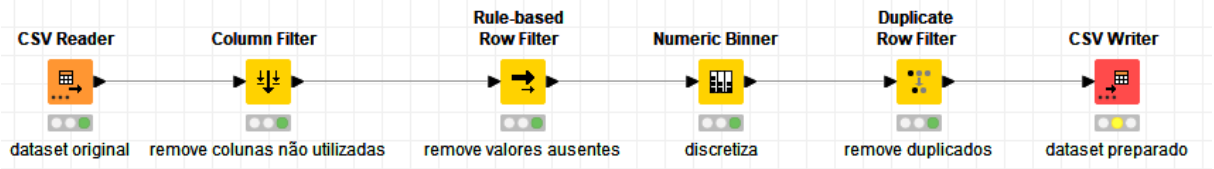
Figura 2 – Discretização do *dataset* Doenças Cardíacas



Fonte: Elaborado pelo autor

DS3) Renda Adulta: contém 32.561 instâncias com 14 atributos (6 contínuos e 8 categóricos) e 2 classes ($<= 50K$ e $> 50K$). Contém informações demográficas e econômicas de indivíduos e é utilizado para prever se uma pessoa ganha mais de 50 mil dólares por ano. É tido como ideal para estudos de discriminação de renda e análise socioeconômica. O atributo [fnlwgt], que é um ponderador estatístico utilizado para ajustar a precisão da amosta com relação à população que ela representa, foi excluído. O atributo education-num foi também excluído do dataset, já que é contínuo e outro atributo, education, é discreto e contém os mesmos rótulos que seriam aplicadas a education-num se discretizado. Foram removidas as 2.399 instâncias que continham valores ausentes, restando 30.162, que tiveram seus atributos contínuos discretizados. Em seguida foi realizada a remoção das instâncias repetidas, restando 12.193 registros. Esse processo de discretização e posterior remoção de registros duplicados foi realizado com a utilização do Knime e o fluxo pode ser visto na Figura 3.

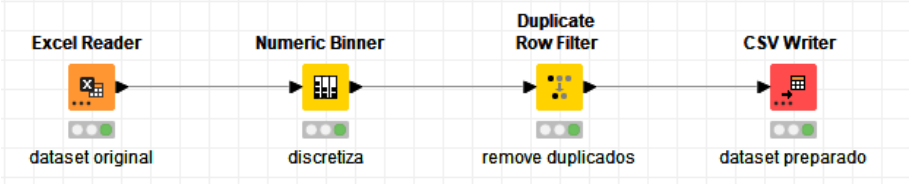
Figura 3 – Discretização do *dataset* Renda Adulta



Fonte: Elaborado pelo autor

DS4) Credit Card Default: contém 30.000 instâncias com 23 atributos (19 contínuos e 4 categóricos) e duas classes (Adimplente, Inadimplente). Um rico conjunto de dados que inclui demografia, histórico de pagamentos, crédito e dados de inadimplência, usado para prever a inadimplência de cartão de crédito. Os atributos contínuos foram discretizados e, em seguida, foi realizada a remoção das instâncias repetidas, restando 16.995 registros. Esse processo de discretização e posterior remoção de registros duplicados foi realizado com a utilização do Knime e o fluxo pode ser visto na Figura 4.

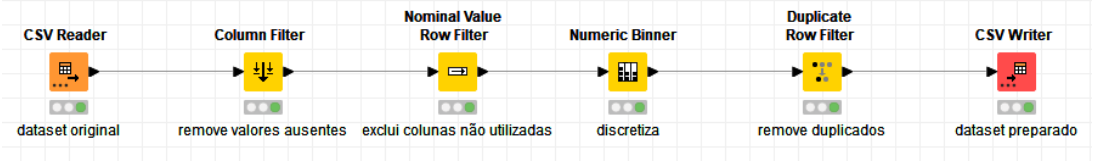
Figura 4 – Discretização do *dataset* Credit Card Default



Fonte: Elaborado pelo autor

DS5) Bank Marketing: contém 45.211 instâncias com 16 atributos (5 contínuos e 11 categóricos) e 2 classes (Sim, Não). Relacionado a campanhas de marketing direto (ligações telefônicas) de uma instituição bancária portuguesa. O objetivo de classificação é prever se o cliente fará um depósito a prazo. Foram removidos os atributos contact e poutcome porque estavam com unknown (desconhecido) em todos os valores. Também foram removidos os atributos day e month porque são apenas o registro de quando o contato telefônico foi realizado. Os atributos contínuos foram discretizados e, em seguida, foi realizada a remoção das instâncias repetidas, restando 9.059 registros. Esse processo de discretização e posterior remoção de registros duplicados foi realizado com a utilização do Knime e o fluxo pode ser visto na Figura 5.

Figura 5 – Discretização do *dataset* Bank Marketing



Fonte: Elaborado pelo autor

5.2 Cálculo dos Índices de Qualidade e da Função Utilidade

Na Tabela 6 é possível observar o resultados da aplicação das medidas de qualidade nas bases de dados. Para os aspectos Rastreabilidade, Integridade, Privacidade, Disponibilidade, Interpretabilidade e Consistência semântica foi atribuído o valor máximo (de 1) para cada medida. Foi feito desta forma para viabilizar o experimento, uma vez que a atribuição de índices para esses aspectos não é feita de forma automatizada, mas sim por inspeção humana. Também para o aspecto Justiça foi atribuído o valor máximo (de 1) porque é preciso a atuação humana na definição de quais aspectos são sensíveis à Justiça. Para a Reprodutibilidade foram extraídas duas amostras extratificadas (com a mesma proporção de classes do dataset completo), contendo cada uma 20% do total de instâncias, para a aplicação do teste de homogeneidade. Para todos os aspectos foi atribuído o mesmo peso W na Função Utilidade, ficando então $W_j = \frac{1}{13} \forall_j$.

Tabela 6 – Resultados da proposta de métricas nos *datasets* do UCI

Aspecto de Qualidade	FU	Dataset 1 - Iris				Dataset 2 - Doenças Cardíacas				Dataset 3 - Renda Adulta				Dataset 4 - Credit Card Default				Dataset 5 - Bank Marketing			
		Índ. Qld.	Vmin	Beta	FU	Índ. Qld.	Vmin	Beta	FU	Índ. Qld.	Vmin	Beta	FU	Índ. Qld.	Vmin	Beta	FU	Índ. Qld.	Vmin	Beta	FU
1) Rastreabilidade	Ex	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000
2) Integridade	Li	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000
3) Privacidade	Li	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000
4) Disponibilidade	Ex	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000
5) Relevância	Li	0,3750	0,5		0,0000	0,6154	0,5		0,2308	0,5833	0,5		0,1666	0,5217	0,5		0,0435	0,5417	0,5		0,0833
6) Interpretabilidade	Ex	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000	1,0000		10	1,0000
7) Consistência: Sintática	Li	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000
Semântica	Li	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000
8) Diversidade	Si	0,2109		10	0,0031	0,0002		10	0,0000	0,0001		10	0,0000	0,0000		10	0,0000	0,0061		10	0,0001
9) Eficácia	Se	1,0000		10	0,0000	2,4896		10	0,0000	4,0053		10	0,0000	5,4555		10	0,0000	1,1378		10	0,0000
10) Justiça	Li	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000	1,0000	0,5		1,0000
11) Representatividade	Si	1,0000		10	1,0000	0,0000		10	0,0000	0,0000		10	0,0000	0,0000		10	0,0000	0,0833		10	0,0002
12) Reprodutibilidade	Ex	1,0000		10	1,0000	1,0000		10	1,0000	0,9167		10	0,9999	0,9130		10	0,9999	1,0000		10	1,0000
Total					**0,7695**				**0,7101**				**0,7051**				**0,6956**				**0,6987**

Fonte: Elaborado pelo autor

5.3 Experimentos Para Verificação da Assertividade da Proposta de Métricas

Nesta seção são feitos experimentos com diferentes manipulações dos *datasets*, a fim de demonstrar a variação correspondente dos índices de qualidade aferidos pela proposta de métricas nos aspectos afetados pela manipulação realizada. Esses experimentos serão realizados com apenas 5 aspectos, que possuem índices aferidos de forma sistemática, permitindo a automatização do seu cálculo. São eles: Relevância, Diversidade, Eficácia, Justiça e Reprodutibilidade. Para os demais índices, esse experimento não faz sentido, pois eles possuem índices atribuídos por avaliação humana.

Nestes experimentos foi aplicada a proposta de métricas, para definir os índices de qualidade e as funções utilidade. Em seguida foram realizados diferentes experimentos com variações do *dataset* pertinentes ao aspecto avaliado. Para cada *dataset* e suas variações também foi construída uma árvore de decisão, que é um algoritmo de aprendizado de máquina de simples entendimento, para ver seus resultados e relacionar suas variações com as variações dos resultados da proposta de métricas em cada experimento. Foram construídas árvores de decisão no Knime que utilizam o índice Gini e deixando somente 2% da base para validação, que é o mínimo permitido pelo Knime, para maximizar a base de aprendizado do algoritmo.

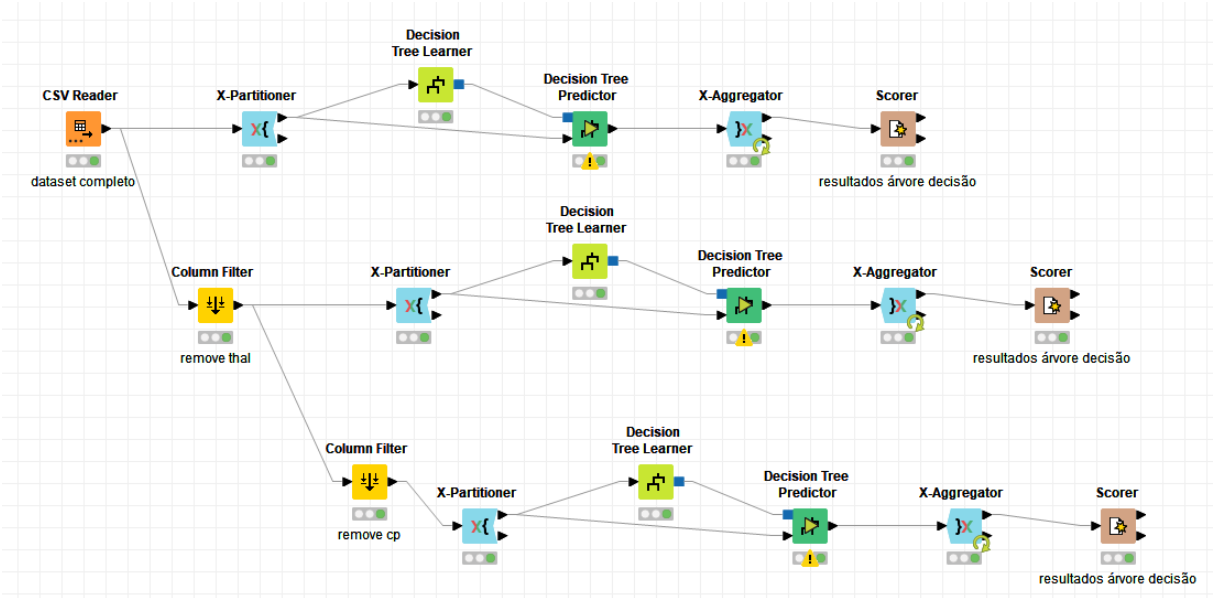
05) Relevância: no primeiro experimento foi removido o atributo de menor Entropia em cada *dataset*, que é: em Doenças Cardíacas o [thal] e em Renda Adulta o [relationship]. No segundo experimento foram removidos os dois atributos de menor Entropia de cada *dataset*, que são: em Doenças Cardíacas [thal e cp] e em Renda Adulta [relationship e marital-status]. Na Tabela 7 podem ser vistos os resultados desses experimentos, que mostram que as mudanças nos *datasets* que provocaram a redução do índice de qualidade e da utilidade aferidos pela proposta de métricas também provocaram a queda da acurácia da árvore de decisão. Nas Figuras 6 e 7 podem ser vistas as árvores de decisão desenvolvidas no Knime.

Tabela 7 – Relevância - Resultados dos Experimentos

Dataset	Dataset Completo			Sem Atributo de Menor Entropia			Sem 2 Atributos de Menores Entropias		
	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia
Doenças Cardíacas	0,6154	0,2308	87,46%	0,5833	0,1666	85,66%	0,5454	0,0908	83,51%
Renda Adulta	0,5833	0,1666	92,40%	0,5454	0,0908	92,22%	0,5000	0,0000	89,69%

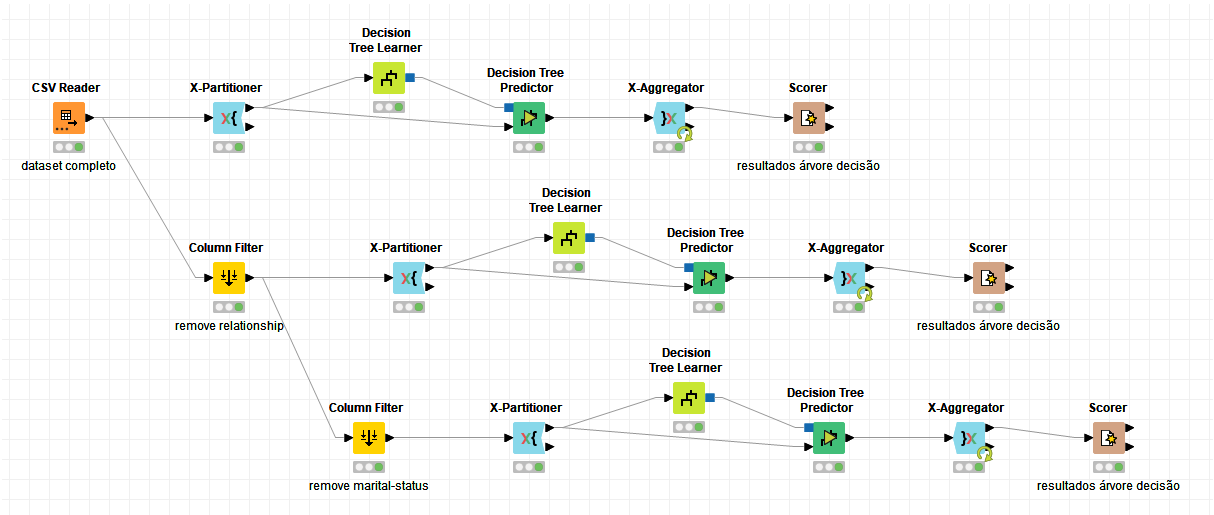
Fonte: Elaborado pelo autor

Figura 6 – Árvore de decisão Relevância - Doenças Cardíacas



Fonte: Elaborado pelo autor

Figura 7 – Árvore de decisão Relevância - Renda Adulta



Fonte: Elaborado pelo autor

08) Diversidade: no primeiro experimento foi removida metade dos registros de forma aleatória. No segundo experimento foram removidos 75% dos registros do *dataset* original de forma aleatória. O processo de geração destes *datasets* de amostras para o experimento

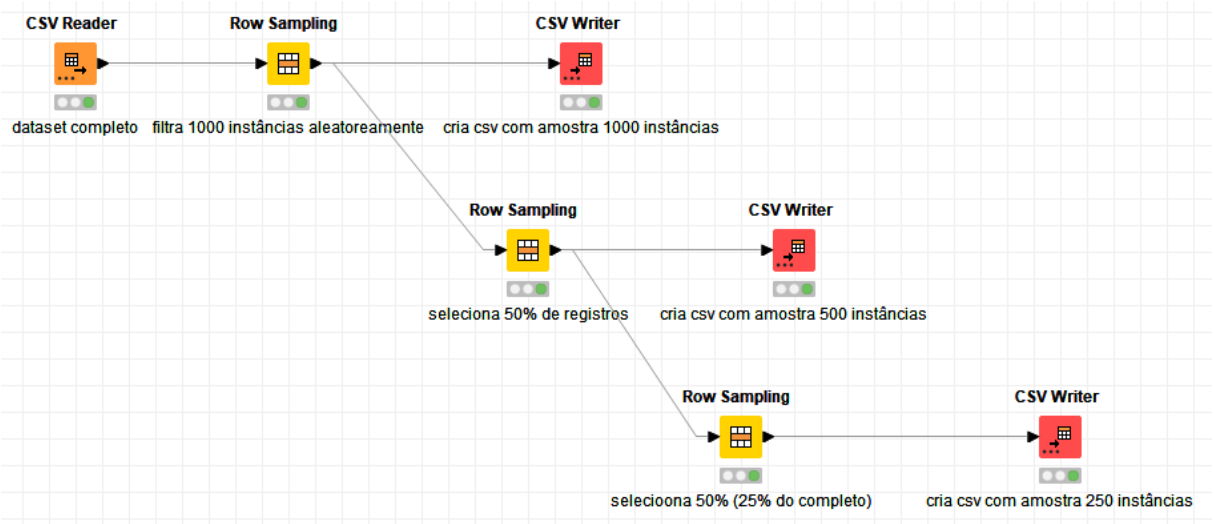
foi realizado com o Knime e na Figura 9 pode ser visto o fluxo criado para essas gerações. Na Tabela 8 podem ser vistos os resultados desses experimentos, que mostram que as mudanças no *dataset* que provocaram a redução na qualidade aferida pela proposta de métricas também provocou, na maioria dos casos, queda da acurácia do algoritmo da árvore de decisão. Como esse aspecto, pela sua forma de cálculo, geralmente resulta em índices de qualidade de valor muito baixo e, com isso, é difícil se ter uma variação na utilidade em experimentos, foi considerada a variação dos índices de qualidade, que já mostra variação no mesmo sentido da variação da acurácia da árvore de decisão. Para o *dataset* Renda Adulta foi realizado um segundo experimento partindo de uma amostra de 1000 instâncias do *dataset* escolhidas aleatoriamente, e neste novo experimento foi possível observar a variação da acurácia da árvore de decisão indo na mesma direção da variação do índice de qualidade visto no experimento com o *dataset* completo. O fluxo de geração dessas amostras no Knime pode ser visto na Figura 8. Na Figura 10 pode ser vista a Árvore de Decisão desenvolvida no Knime.

Tabela 8 – Diversidade - Resultados dos Experimentos

Dataset	Dataset Completo			50% das Classes do Dataset			25% das Classes do Dataset		
	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia
Doenças Cardíacas	0,000249	0,0000	74,55%	0,000125	0,0000	72,66%	0,000063	0,0000	71,01%
Renda Adulta	0,000008	0,0000	82,78%	0,000004	0,0000	82,08%	0,000002	0,0000	82,49%
Renda Adulta (1000 inst.)	0,000001	0,0000	82,93%	0,000000	0,0000	79,87%	0,000000	0,0000	76,75%

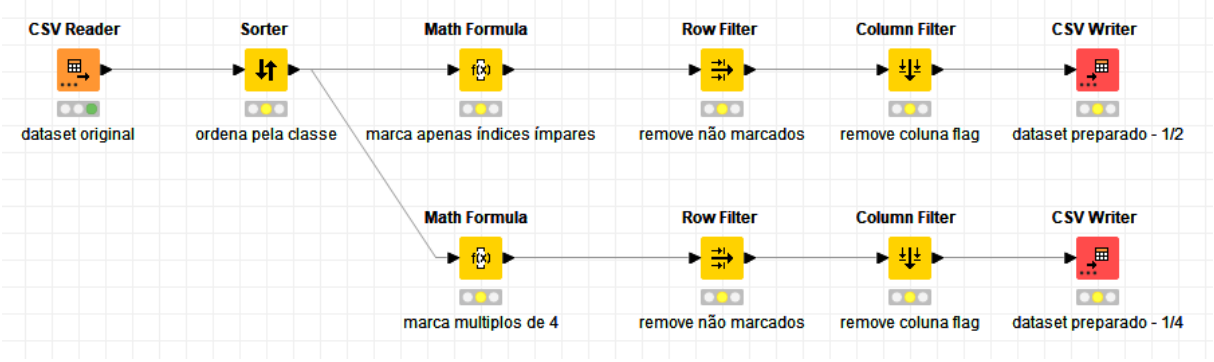
Fonte: Elaborado pelo autor

Figura 8 – Seleção de 1000 registros da Renda Adulta



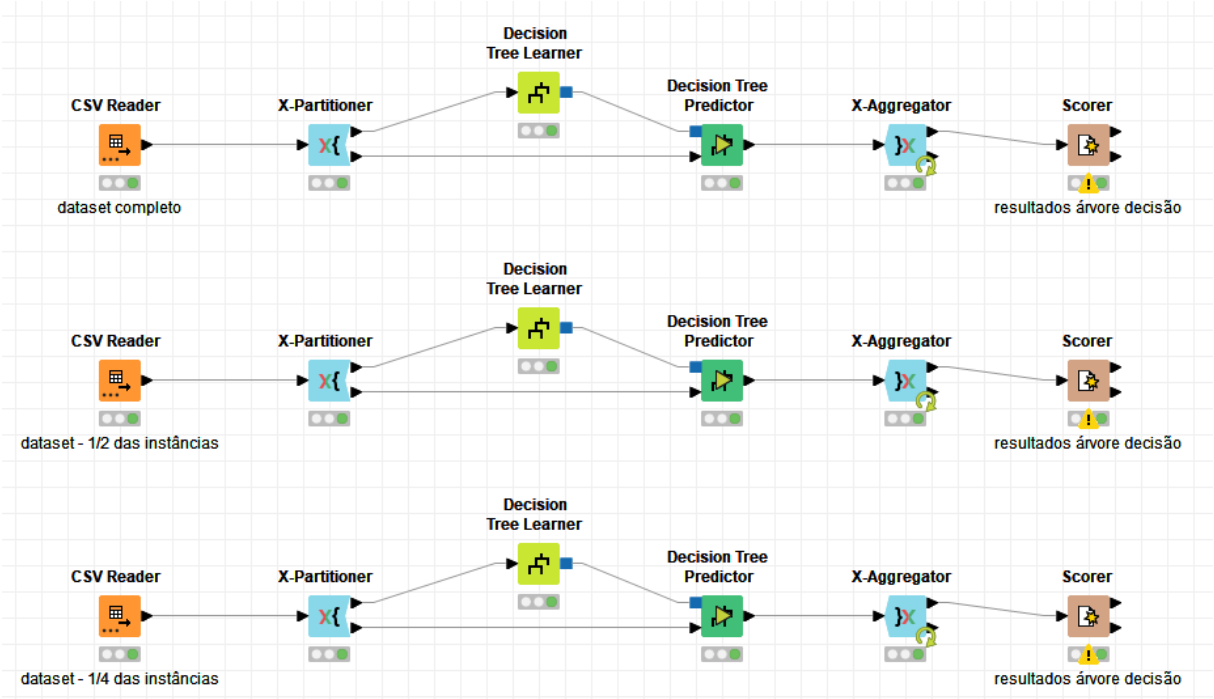
Fonte: Elaborado pelo autor

Figura 9 – Diversidade - Geração de Amostras



Fonte: Elaborado pelo autor

Figura 10 – Diversidade - Árvore de decisão



Fonte: Elaborado pelo autor

09) Eficácia: no primeiro experimento foram removidos os 6 atributos de menor Entropia em cada *dataset*, que são: em Doenças Cardíacas [thal, cp, ca, oldpeak, exang e thalach] e em Renda Adulta o [relationship, marital-status, education, occupation, capital-gain e hours-per-week]. No segundo experimento foram removidos, além dos anteriores, mais 3 atributos de menor Entropia de cada *dataset*, que são: em Doenças Cardíacas [slope, sex e age] e em Renda Adulta [workclass, sex e native-country]. Na Tabela 9 podem ser vistos os resultados desses experimentos, que mostram que quanto menor o índice de qualidade, maior a utilidade desse aspecto e, nesse caso, menor a acurácia da árvore de decisão, o que é esperado porque estamos reduzindo a quantidade de informação do universo estudado contida no *dataset* e, com isso, a árvore de decisão tem

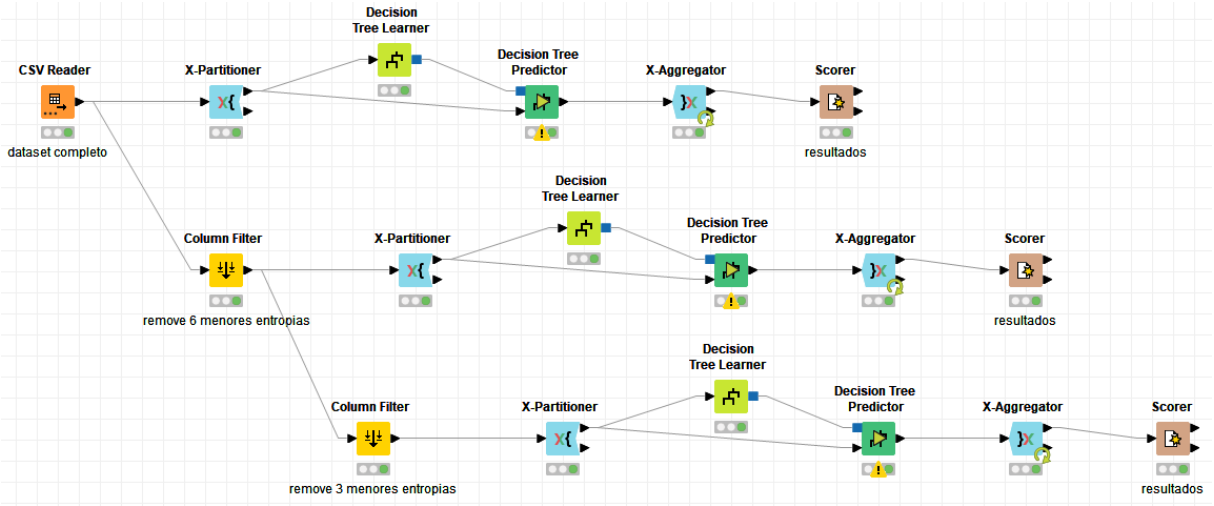
menor eficácia. É importante ressaltar que é exatamente por esse crescimento inverso da qualidade com o aprendizado do dataset que esse aspecto utiliza a função de utilidade sigmoidal invertida. Na Figura 11 pode ser vista a Árvore de Decisão desenvolvida no Knime.

Tabela 9 – Eficácia - Resultados dos Experimentos

Dataset	Dataset Completo			Sem 6 Atributos de Menores Entropias			Sem 9 Atributos de Menores Entropias		
	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia
Doenças Cardíacas	2,4896	0,1203	87,10%	3,1002	0,0691	77,42%	3,9156	0,0318	61,65%
Renda Adulta	4,0053	0,0291	92,50%	5,5303	0,0065	80,17%	7,2785	0,0011	78,59%

Fonte: Elaborado pelo autor

Figura 11 – Eficácia - Árvore de decisão



Fonte: Elaborado pelo autor

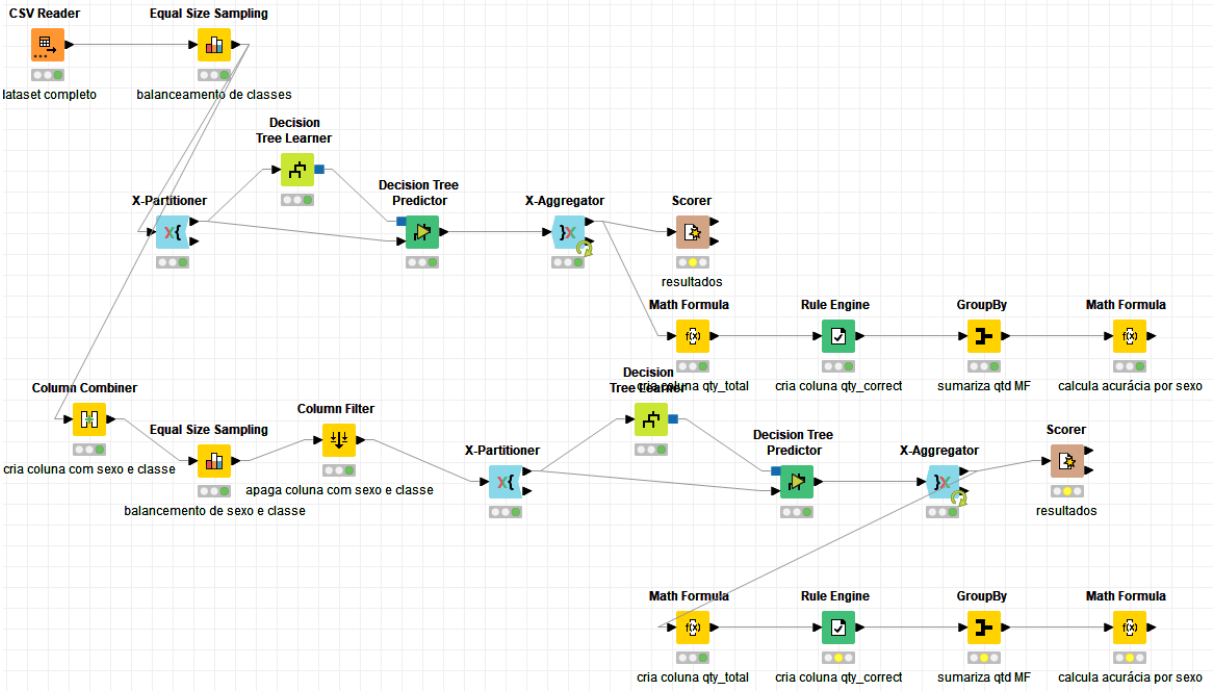
10) Justiça: realizado o experimento com o atributo Sexo, que é sensível à Justiça e está presente nos três *datasets* utilizados. No experimento os *datasets* foram balanceados pelo Sexo e pela classe simultaneamente. O método utilizado nos balanceamentos foi o *undersampling*. Na Tabela 10 podem ser vistos os resultados desses experimentos, que mostram que o balanceamento que provocou o aumento do índice de qualidade e da utilidade dos datasets também provocou a redução da diferença de acurácia entre os sexos na árvore de decisão. Na Figura 12 pode ser vista a Árvore de Decisão desenvolvida no Knime.

Tabela 10 – Justiça - Resultados dos Experimentos

Dataset	Tipo Amostra	Quantidade de Instâncias				Índice de Qualidade	Função Utilidade	Árvore de Decisão - Acurácia			
		Sex = F	Sex = M	Class = F	Class = V			Total	Sex = F	Sex = M	Diff (MxF)
Doenças Cardíacas	Completa	96	201	160	137	0,0000	0,0000	93,60%	95,83%	92,54%	3,30%
	Balanceada - sex + class	50	50	50	50	1,0000	1,0000	89,00%	88,00%	90,00%	2,00%
Renda Adulta	Completa	9782	20380	22654	7508	0,0000	0,0000	93,73%	95,39%	92,64%	2,75%
	Balanceada - sex + class	2224	2224	2224	2224	1,0000	1,0000	90,87%	91,81%	89,92%	1,89%
Credit Card Default	Completa	18112	11888	6636	23364	0,0000	0,0000	90,92%	91,17%	90,57%	0,60%
	Balanceada - sex + class	5050	5050	5050	5050	1,0000	1,0000	88,87%	89,02%	88,72%	0,30%

Fonte: Elaborado pelo autor

Figura 12 – Justiça - Árvore de decisão



Fonte: Elaborado pelo autor

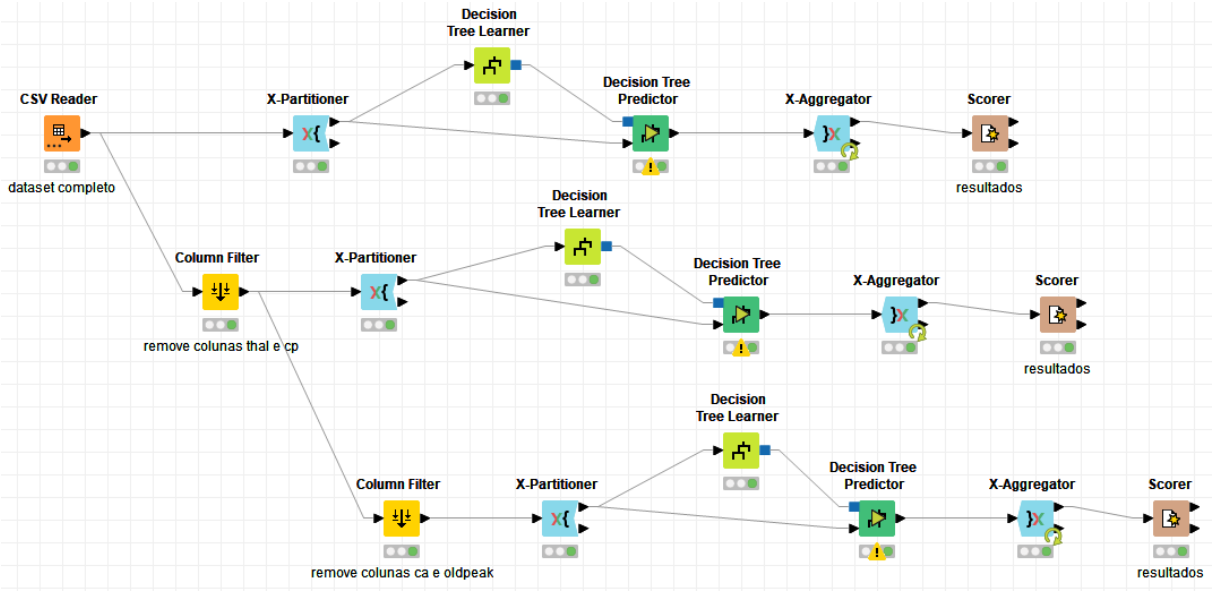
12) Reprodutibilidade: no primeiro experimento foram removidos os dois atributos de menor Entropia em cada *dataset*, que são: em Doenças Cardíacas [thal e cp] e em Renda Adulta [relationship e marital-status]. No segundo experimento foram removidos os quatro atributos de menor Entropia de cada *dataset*, que são: em Doenças Cardíacas [thal, cp, ca e oldpeak] e em Renda Adulta [relationship, marital-status, education e occupation]. Em ambos os experimentos foi utilizada a mesma metodologia aplicada à base completa para cálculo da qualidade, que foi a extração de duas amostras extratificadas contendo cada uma 20% do total de instâncias do dataset completo para a aplicação do teste de homogeneidade. Na Tabela 11 podem ser vistos os resultados desses experimentos, onde pode ser comprovado que as manipulações do *dataset* que provocaram a queda da qualidade aferida e da utilidade resultante também provocaram a redução da acurácia da árvore de decisão. Nas Figuras 13 e 14 podem ser vistas as árvores de decisão desenvolvidas no Knime.

Tabela 11 – Reprodutibilidade - Resultados dos Experimentos

Dataset	Dataset Completo			Sem 2 Atributos de Menores Entropias			Sem 4 Atributos de Menores Entropias		
	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia	Índice Qualidade	Função Utilidade	Árvore Decisão Acurácia
Doenças Cardíacas	1,0000	1,0000	88,17%	0,9167	0,9999	84,95%	0,8818	0,9999	79,21%
Renda Adulta	0,9167	0,9999	92,48%	0,8182	0,9997	89,43%	0,7000	0,9991	83,57%

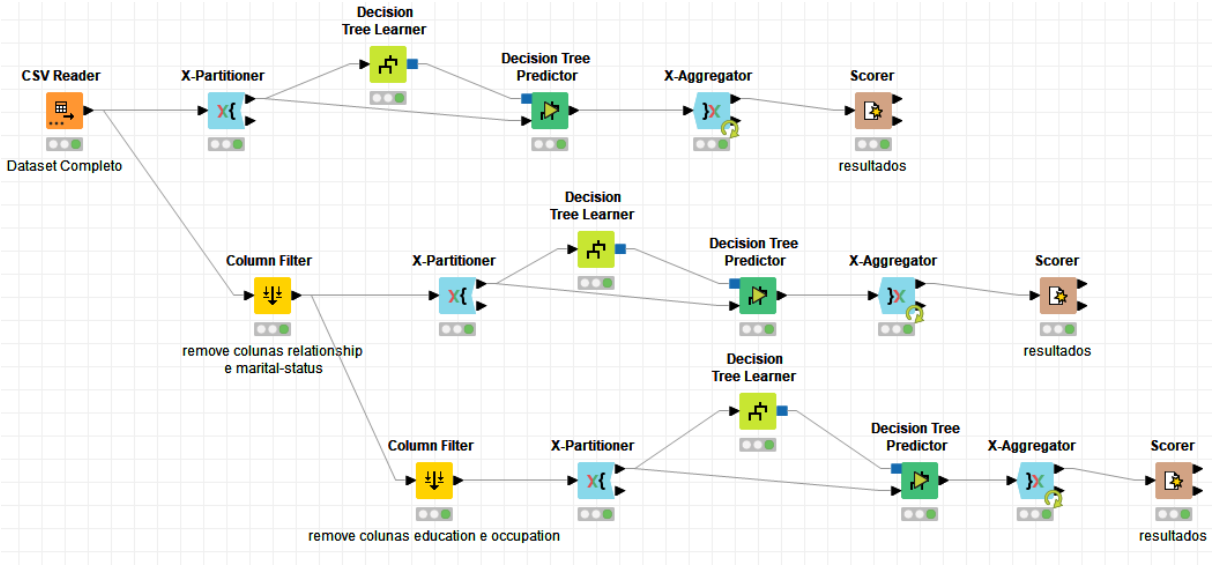
Fonte: Elaborado pelo autor

Figura 13 – Reprodutibilidade - Árvore de decisão - Doenças Cardíacas



Fonte: Elaborado pelo autor

Figura 14 – Reprodutibilidade - Árvore de decisão - Renda Adulta



Fonte: Elaborado pelo autor

6 CONCLUSÕES

Esse trabalho apresentou uma revisão da literatura que identifica os principais aspectos de qualidade que são apontados como sendo os mais importantes a serem avaliados em *datasets* que serão utilizados em projetos de aprendizado de máquina, de forma a conhecer previamente a contribuição que esses dados podem dar para o aprendizado desejado no projeto. Foram identificados diversos aspectos, dos quais foram escolhidos 12 para a elaboração de uma proposta de métricas de qualidade que tem o propósito de identificar o quanto o *dataset* pode contribuir com o objetivo do aprendizado buscado pelo projeto. Isso melhoraria a visão prévia das possibilidades de sucesso do projeto, além de mostrar quais ajustes nos dados podem contribuir para a melhora dos resultados que serão obtidos.

Foi demonstrada a aplicação das métricas propostas em um conjunto de dados toy, para exemplificar a sua forma de utilização. Também foram escolhidos 5 *datasets* do UCI para aplicação das métricas, a fim de demonstrar sua aplicação em cenários reais, e os resultados dessa aplicação foram demonstrados no trabalho. Para os aspectos cuja proposta de métrica pode ser sistematizada, permitindo cálculo automatizado para qualquer *dataset*, foram realizados experimentos com alguns dos *datasets* do UCI utilizados na aplicação citada anteriormente, a fim de comprovar a assertividade dos índices aferidos pela proposta de métricas. Após todos os trabalhos de definição, aplicação e experimentação das métricas propostas com *datasets* toy e reais, pode-se dizer dos aspectos e da proposta de métricas de avaliação qualidade:

- **Rastreabilidade:** aspecto útil principalmente quando os dados precisam passar por diversos processos de cópia ou transformações até sua disponibilização para o projeto. Na proposta não foi apresentada uma forma automatizada de cálculo desse índice de qualidade, mas na aplicação em cada projeto é possível que haja alguma ferramenta ou conjunto de ferramentas que possibilite essa automatização.
- **Integridade:** é um aspecto muito importante de ser avaliado principalmente quando a geração e disponibilização dos dados passa por diversos locais e ferramentas. O artifício proposto, que utiliza de *hashes* para validar essa integridade, é um facilitador da aplicação da métrica, pois existem diversas formas, muitas delas bem simples, de calcular esses *hashes*.
- **Privacidade:** esse é um aspecto cuja avaliação tem se tornado cada vez mais importante, considerando que é crescente a preocupação com a exposição de dados

pessoais e já existem diversas leis, inclusive, que tratam de cuidados que é preciso ter com a privacidade de diversos dados. No Brasil temos a Lei Geral de Proteção de Dados Pessoais (LGPD), como exemplo, que especifica regras para manutenção da privacidade dos dados e prevê punições para projetos que não respeitarem essas regras. Então, esse aspecto de qualidade torna-se um recurso poderoso na identificação de possíveis falhas no cumprimento dessas regras, possibilitando os ajustes e adequações já no início dos projetos.

- **Disponibilidade:** a avaliação desse aspecto como foi proposta, ainda que feita de forma humana, pode ajudar a identificar, já na concepção do projeto, atributos que precisarão ser eliminados ou substituídos para não comprometer a conclusão do projeto ou o cumprimento dos prazos estabelecidos.

- **Relevância:** a forma de métrica proposta ajuda a escolher atributos que influenciam diretamente na qualidade do aprendizado obtido com o *dataset* no projeto de aprendizado de máquina. Isso pode facilitar o alinhamento das expectativas referentes aos resultados do projeto, permitindo ações já no início que possibilitem o aumento dessas expectativas.

- **Interpretabilidade:** é um aspecto que não se aplica a todos os tipos de projeto de Aprendizado de Máquina, mas nos que se aplica pode apontar previamente a possível ocorrência de problemas de compreensão do aprendizado que será obtido pelo projeto, podendo até mesmo apontar uma eventual inviabilidade do projeto com o dataset disponível.

- **Consistência - Sintática:** esse aspecto e sua forma de métrica são bastante simples de compreender e aplicar, além de serem extremamente úteis na grande maioria dos projetos, pois identifica e permite a eliminação prévia de valores inválidos que poderiam comprometer o aprendizado do projeto.

- **Consistência - Semântica:** mesmo exigindo um trabalho humano cuja complexidade pode ser grande em *datasets* de muitos atributos correlacionados, a avaliação desse aspecto, assim como o de Consistência Sintática, pode permitir a identificação prévia de valores inválidos, possibilitando o acerto ou eliminação dos registros que poderiam comprometer o aprendizado do projeto antes de sua construção.

- **Diversidade:** mesmo sendo um aspecto que tem como resultado da sua forma de cálculo proposta um índice de qualidade que na maioria das amostras ficará muito baixo e resultará em uma utilidade 0 ou próximo disso, ainda assim pode ser muito útil porque mostra a distância do tamanho da amostra contida no *dataset* para o tamanho do universo estudado. Isso pode apontar previamente a necessidade de se obter mais amostras a fim de viabilizar o sucesso do projeto.

- **Eficácia:** um dos grandes benefícios que a avaliação desse aspecto pode trazer é a otimização de recursos em projetos onde a quantidade de recursos consumidos é diretamente influenciada pela quantidade de atributos do *dataset* utilizado. Com esse aspecto é possível que se avalie previamente a redução dos atributos que serão utilizados no Aprendizado de Máquina sem que isso reduza de forma significativa a qualidade dos resultados obtidos.

- **Justiça:** a avaliação desse aspecto pode ser muito útil por possibilitar, já na fase de preparação do *dataset* para o projeto, a identificação da necessidade de pré-processamentos que impedirão que os resultados do aprendizado sejam injustamente tendenciosos. Com o momento em que pode ser vista uma explosão do uso de Aprendizado de Máquina para geração de conhecimento que impacta diretamente na vida das pessoas, a criação de modelos injustos é uma preocupação muito latente e, com isso, esse aspecto se torna extremamente importante para muitos projetos.

- **Representatividade:** assim como na Justiça, a avaliação desse aspecto pode identificar a necessidade de tratamentos no *dataset* que elimine tendências no aprendizado, mas nesse caso de forma mais abrangente, compreendendo todos os tipos de atributos, tanto de pessoas quanto de coisas e situações. Então, é bastante importante na maioria dos projetos para eliminar tendências que podem acabar reduzindo a utilidade dos resultados obtidos.

- **Reprodutibilidade:** esse aspecto pode ser bastante útil, principalmente onde amostras são difíceis de se conseguir ou então são muito pequenas a ponto de dificultar o aprendizado desejado. Como aspecto poderá indicar uma eventual facilidade de se gerar novas bases de aprendizado a partir das amostras já obtidas, nesses casos isso pode apontar a possibilidade de enriquecimento da base de aprendizado e, com isso, a melhora dos resultado obtidos.

Como possibilidade de evolução futura desse trabalho é importante aumentar o volume de testes. Isso pode ser feito com a realização de mais experimentos utilizando os mesmos *datasets* e também com a adição de outros *datasets*. Também pode ser apontado o estudo da possibilidade de cálculos automatizados das métricas dessa proposta que ainda dependem de intervenção humana, pois isso facilitaria sua aplicação.

Como uma evolução visando o aumento do escopo de atuação da proposta também podem ser estudadas métricas para atributos contínuos, pois isso poderia permitir a utilização de atributos que não podem ser categorizados. E mesmo para os que podem, isso poderia facilitar a aplicação da proposta por não exigir a categorização desses atributos, o que também possibilitaria a utilização dos seus valores com maior precisão.

Além disso a proposta também pode ser expandida para contemplar mais aspectos apontados pela literatura, de forma a torná-la mais abrangente e com possibilidade de utilização mais ampla. Também pode ser estudada a construção de um *framework* dessa proposta para cálculo automatizado da qualidade dos *datasets*. Esse estudo poderia resultar em uma ferramenta de uso simples e rápido por qualquer pessoa, mesmo que não tenha o entendimento do cálculo das métricas de cada aspecto de qualidade, bastando apenas conhecer a definição e aplicabilidade de cada um para utilizar a metodologia de forma eficaz.

REFERÊNCIAS

ALAOUI, I. E.; GAHI, Y.; MESSOUSSI, R. Big data quality metrics for sentiment analysis approaches. BBDE 2019: Proceedings of the 2019 International Conference on Big Data Engineering, p. 36–43, jan. 2019. ISSN 0950-5849. Disponível em: <https://doi.org/10.1145/3341620.3341629>.

ARASS, M. E.; SOUISSI, N. Data lifecycle: From big data to smartdata. In: 2018 IEEE 5th International Congress on Information Science and Technology (CiSt). [S.l.: s.n.], 2018. p. 80–87.

ARASS, M. E.; TIKITO, I.; SOUISSI, N. Data lifecycles analysis: Towards intelligent cycle. In: 2017 Intelligent Systems and Computer Vision (ISCV). [S.l.: s.n.], 2017. p. 1–8.

BATINI, C. et al. Methodologies for data quality assessment and improvement. ACM Computing Surveys, p. 1–52, July 2009. ISSN 0950-5849. Disponível em: <https://dl.acm.org/doi/10.1145/1541880.1541883>.

BEIL, M. et al. Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Medicine Experimental, v. 7, p. 70, 2019. ISSN 2197-425X. Disponível em: <https://doi.org/10.1186/s40635-019-0286-6>.

BERTOSSI, L.; GEERTS, F. Data quality and explainable ai. J. Data and Information Quality, Association for Computing Machinery, New York, NY, USA, v. 12, n. 2, may 2020. ISSN 1936-1955. Disponível em: <https://doi.org/10.1145/3386687>.

BLAKE, R.; MANGIAMELI, P. The effects and interactions of data quality and problem complexity on classification. J. Data and Information Quality, Association for Computing Machinery, New York, NY, USA, v. 2, n. 2, feb 2011. ISSN 1936-1955. Disponível em: <https://doi.org/10.1145/1891879.1891881>.

BOYD, K. L. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. Proc. ACM Hum.-Comput. Interact., Association for Computing Machinery, New York, NY, USA, v. 5, n. CSCW2, oct 2021. Disponível em: <https://doi.org/10.1145/3479582>.

BRADY, A. P.; NERI, E. Artificial intelligence in radiology—ethical considerations. Diagnostics, v. 10, n. 4, 2020. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/10/4/231>.

CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. In: Data Science Journal, vol. 14, no. 0. [S.l.: s.n.], 2015. p. 1–10.

CARTER, S. M. et al. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Nationa Libray of Medicine, Elsevier, v. 49, p. 25–32, 2020. ISSN 0960-9776.

CHAR, D. S.; ABRàMOFF, M. D.; FEUDTNER, C. Identifying ethical considerations for machine learning healthcare applications. The American Journal of Bioethics, Taylor Francis, v. 20, n. 11, p. 7–17, 2020. PMID: 33103967.

CHATILA, R. et al. The ieee global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. IEEE Robotics Automation Magazine, v. 24, n. 1, p. 110–110, 2017.

CHEN, H.; CHEN, J.; DING, J. Data evaluation and enhancement for quality improvement of machine learning. IEEE Transactions on Reliability, v. 70, n. 2, p. 831–847, 2021.

CHEN, I. Y. et al. Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, v. 4, n. 1, p. 123–144, 2021.

CLARKE, R. Principles and business processes for responsible ai. Computer Law Security Review, v. 35, n. 4, p. 410–422, 2019. ISSN 0267-3649. Disponível em: <https://www.sciencedirect.com/science/article/pii/S026736491930127X>.

CURRIE, G.; HAWK, K. E. Ethical and legal challenges of artificial intelligence in nuclear medicine. Seminars in Nuclear Medicine, v. 51, n. 2, p. 120–125, 2021. ISSN 0001-2998. Artificial Intelligence in Nuclear Medicine. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0001299820300921>.

DING, J. et al. A case study of the augmentation and evaluation of training data for deep learning. J. Data and Information Quality, Association for Computing Machinery, New York, NY, USA, v. 11, n. 4, aug 2019. ISSN 1936-1955. Disponível em: <https://doi.org/10.1145/3317573>.

HE, T. et al. From data quality to model quality: An exploratory study on deep learning. Internetware '19: Proceedings of the 11th Asia-Pacific Symposium on Internetware, p. 1–6, out. 2019. ISSN 0950-5849. Disponível em: <https://doi.org/10.1145/3361242.3361260>.

ISO/IEC 25010. ISO/IEC 25010:2011, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. 2011.

JIAN, G. Artificial intelligence in healthcare and medicine: Promises, ethical challenges and governance. Chinese Medical Sciences Journal, Chinese Medical Sciences Journal, v. 34, n. 2, p. 76, 2019. Disponível em: <http://cmsj.cams.cn/EN/abstract/article$_2$858.*shtml*>.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering - a systematic literature review. Inf. Softw. Technol., Butterworth-Heinemann, USA, v. 51, n. 1, p. 7–15, jan. 2009. ISSN 0950-5849. Disponível em: <https://doi.org/10.1016/j.infsof.2008.09.009>.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. Artificial Intelligence, v. 97, n. 1, p. 273–324, 1997. ISSN 0004-3702. Relevance. Disponível em: <https://www.sciencedirect.com/science/article/pii/S000437029700043X>.

LAM, K. et al. Investigating the ethical and data governance issues of artificial intelligence in surgery: Protocol for a delphi study. JMIR Res Protoc, v. 10, n. 2, p. e26552, Feb 2021. ISSN 1929-0748. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/33616543>.

LANDAU, A. Y. et al. Developing machine learning-based models to help identify child abuse and neglect: key ethical challenges and recommended solutions. Journal of the American Medical Informatics Association, v. 29, n. 3, p. 576–580, 01 2022. ISSN 1527-974X. Disponível em: <https://doi.org/10.1093/jamia/ocab286>.

MARTINEZ-MARTIN, N. et al. Data mining for health: staking out the ethical territory of digital phenotyping. npj Digital Medicine, v. 1, p. 68, 2018. ISSN 2398-6352. Disponível em: <https://doi.org/10.1038/s41746-018-0075-8>.

MCCRADDEN, M. D. et al. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. Canadian Medical Association Open Access Journal, Canadian Medical Association Open Access Journal, v. 8, n. 1, p. E90–E95, 2020. Disponível em: <https://www.cmajopen.ca/content/8/1/E90>.

MICELI, M.; POSADA, J.; YANG, T. Studying up machine learning data: Why talk about bias when we mean power? Proc. ACM Hum.-Comput. Interact., Association for Computing Machinery, New York, NY, USA, v. 6, n. GROUP, jan 2022. Disponível em: <https://doi.org/10.1145/3492853>.

MULLINS, M.; HOLLAND, C. P.; CUNNEEN, M. Creating ethics guidelines for artificial intelligence and big data analytics customers: The case of the consumer european insurance market. Patterns, Elsevier, v. 2, 2021. ISSN 2666-3899. Disponível em: <https://doi.org/10.1016/j.patter.2021.100362>.

MöLLMANN, N. R.; MIRBABAIE, M.; STIEGLITZ, S. Is it alright to use artificial intelligence in digital health? a systematic literature review on ethical considerations. Health Informatics Journal, v. 27, n. 4, p. 14604582211052391, 2021. PMID: 34935557. Disponível em: <https://doi.org/10.1177/14604582211052391>.

PENG, R. D.; MATSUI, E. The Art of Data Science - A Guide for Anyone Who Works with Data. Baltimore, MD, USA: Leanpub, 2018.

PESAPANE, F. et al. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in europe and the united states. Insights into Imaging, v. 9, p. 745–753, 2018. ISSN 1869-4101. Disponível em: <https://doi.org/10.1007/s13244-018-0645-y>.

ROSEMANN, A.; ZHANG, X. Exploring the social, ethical, legal, and responsibility dimensions of artificial intelligence for health – a new column in intelligent medicine. Intelligent Medicine, v. 2, n. 2, p. 103–109, 2022. ISSN 2667-1026. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2667102621001212>.

RUDRARAJU, N. V.; BOYANAPALLY, V. Data Quality Model for Machine learning. 2019. Tese (Doutorado) — Faculty of Computing - Blekinge Institute of Technology - Sweden, Disponível em: <http://urn.kb.se/resolve?urn=urn:nbn:se:bth-18498>.

SAHEB, T.; SAHEB, T.; CARPENTER, D. O. Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis. Computers in Biology and Medicine, v. 135, p. 104660, 2021. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482521004546>.

SCHELTER, S. et al. Automating large-scale data quality verification. Proc. VLDB Endow., VLDB Endowment, v. 11, n. 12, p. 1781–1794, aug 2018. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/3229863.3229867>.

THIRUMURUGANATHAN, S. et al. Automated annotations for ai data and model transparency. J. Data and Information Quality, Association for Computing Machinery, New York, NY, USA, v. 14, n. 1, dec 2021. ISSN 1936-1955. Disponível em: <https://doi.org/10.1145/3460000>.

UCI. UCI. 2025. `https://archive.ics.uci.edu/`.

UNGER, E.; HARN, L.; KUMAR, V. Entropy as a measure of database information. In: [1990] Proceedings of the Sixth Annual Computer Security Applications Conference. [S.l.: s.n.], 1990. p. 80–87.

VAYENA, E.; BLASIMME, A.; COHEN, I. G. Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, Public Library of Science, v. 15, n. 11, p. 1–4, 11 2018.

WINFIELD, A. F. T.; JIROTKA, M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, v. 376, n. 2133, p. 20180085, 2018. Disponível em: <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2018.0085>.