PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EFEITOS DA ANONIMIZAÇÃO NOS PROCESSOS DE MINERAÇÃO DOS DADOS

FELIPE SILVA PAULA

GOIÂNIA
2021

FELIPE SILVA PAULA

EFEITOS DA ANONIMIZAÇÃO NOS PROCESSOS DE MINERAÇÃO DOS DADOS

Trabalho de Conclusão de Curso apresentado à Escola Politécnica, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Sibelius Lellis Vieira
Banca examinadora:
Prof. Me. Gustavo Siqueira Vinhal
Prof. Me. Rafael Leal Martins

GOIÂNIA
2021

FELIPE SILVA PAULA

EFEITOS DA ANONIMIZAÇÃO NOS PROCESSOS DE MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso aprovado em sua forma final pela Escola Politécnica, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Ciência da Computação, em ____/____/_____.

_____

Orientador: Prof. Dr. Sibelius Lellis Vieira

_____

Prof. Me. Gustavo Siqueira Vinhal

_____

Prof. Me. Rafael Leal Martins

_____

Profª. Ma. Ludmilla Reis Pinheiro dos Santos
Coordenadora de Trabalho de Conclusão de Curso

GOIÂNIA

2021

AGRADECIMENTOS

Ao Deus dos exércitos, por me proporcionar forças e sabedoria para passar pelas dificuldades encontradas no decorrer deste trabalho.

Agradeço ao meu orientador Dr. Sibelius Lellis Vieira por aceitar conduzir este trabalho. Sempre paciente e atencioso durante todo este período. Suas ideias e orientações proporcionaram uma experiência inspiradora para mim e promoveram uma imensa diferença no resultado deste trabalho.

Aos meus pais, por todo esforço investido na minha educação. Suas motivações e apoio foram muito importantes nessa minha trajetória, na qual encontrei muitas dificuldades que foram superadas graças aos seus incentivos.

À minha namorada Glenda pela compreensão, paciência e as palavras de incentivo.

RESUMO

Este trabalho propõe-se a aplicar técnicas de anonimização e mineração de dados em um conjunto de dados de crédito, com propósito de verificar a privacidade dos dados, segurança da informação e o cumprimento da Lei Geral de Proteção de Dados Pessoais (LGPD), além da viabilidade de continuar aplicando processos de tomada de decisão nos dados anonimizados. A mineração de dados apresenta-se como uma estratégia para auxiliar na tomada de decisões dos conjuntos de dados anonimizados. Uma pesquisa bibliográfica sobre a LGPD, anonimização e mineração dos dados foi realizada com objetivo de compreender e conceituar estes temas. Com a utilização do *software* Amnesia foram aplicadas as técnicas de anonimização no conjunto de dados de crédito (conjunto original), pois a anonimização pode ser utilizada com intuito de obter a privacidade e o cumprimento de leis e regulamentações de proteção de dados. Essa estratégia pode ser aplicada de modo que não permita a reidentificação dos dados e ainda contribuir com a segurança da informação e gestão de riscos. Para cada uma das técnicas de anonimização aplicadas, a mineração de dados foi empregada com auxílio da ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) e os valores das acurácias obtidos. Como resultados, verifica-se que a anonimização baseada em supressão e generalização garantiu a privacidade necessária, bem como não impediu que técnicas de classificação baseadas em árvores de decisão pudessem ser usadas para tomada de decisão de forma adequada.

Palavras-chave: Anonimização dos dados. Mineração de dados. Lei Geral de Proteção de Dados. Privacidade dos dados. Árvore de decisão.

ABSTRACT

This work proposes to apply anonymization and data mining techniques to a set of credit data, to verify data privacy, information security, and compliance with the General Data Protection Law (GDPL), in addition to the feasibility of continuing to apply decision-making processes to anonymized data. Data mining presents itself as a strategy to assist in decision-making on anonymized datasets. Bibliographical research on the GDPL, anonymization, and data mining was carried out to understand and conceptualize these themes. With the use of the amnesia software, anonymization techniques were applied to the credit dataset (original set), as data anonymization can be used to obtain privacy and compliance with protection laws and regulations of data. This strategy can be applied in a way that does not allow the re-identification of data and still contributes to information security and risk management. For each of the applied anonymization techniques, data mining was used with the help of the Waikato Environment for Knowledge Analysis (WEKA) tool and the accuracy values obtained. As a result, it appears that anonymization based on suppression and generalization ensured the necessary privacy, as well as did not prevent classification techniques based on decision trees from being used for adequate decision making.

Keywords: Data anonymization. Data mining. General data protection law. Data privacy. Decision tree.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ANPD	Autoridade Nacional de Proteção de Dados
ARFF	*Atribute Relation File Format*
CEP	Código de Endereçamento Postal
CPF	Cadastro de Pessoas Físicas
CSV	*Comma Separated Values*
GDPR	*General Data Protection Regulation*
GPDP	Gabinete para Proteção de Dados Pessoais
GT	Grupo de Trabalho do Artigo 29
ISO/IEC	Organização Internacional de Padronização/*International Electrotechnical Commission*
JDBC	*Java Database Connectivity*
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
LGPD	Lei Geral de Proteção de Dados
NBR	Normas Brasileiras
PSI	Política de Segurança da Informação
RG	Registro Geral
RIPD	Relátorio de Impacto a Proteção de Dados Pessoais
UE	União Europeia
URL	*Uniform Resource Locator*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A partir do avanço tecnológico e a crescente evolução da *internet,* serviços baseados em sistemas *online* e *offline* têm sido utilizados para o tratamento de dados e tais informações envolvem dados pessoais que devem ser protegidas contra vazamentos e ataques internos e externos. No decorrer dos anos pode-se observar a grande quantidade de notícias relacionadas a vazamento de dados nas empresas do mundo todo. (PANEK, 2019).

Casos impactantes de vazamento de dados envolvendo grandes empresas como Netflix, Facebook, Uber, Adobe e Netshoes deixaram expostos arquivos com milhões de dados pessoais como nomes, senhas, e-mails, número de celular, Cadastro de Pessoas Físicas (CPF) e entre outros. Nestes arquivos os dados são selecionados e organizados por hackers a fim de disponibilizá-los na *dark web,* local em que informações ilegais podem ser compartilhadas. (DIGITAL, 2018).

Conforme exposto por Cyber (2021), no primeiro semestre de 2020, o aumento nas tentativas de ataques *hackers* realizados em indústrias e empresas foram de 460% voltando a disparar no segundo semestre, em que as tentativas de invasões tiveram um aumento de 860%.

Por outro lado, a Lei Geral de Proteção de Dados (LGPD) se apresenta com propósito de proteger os direitos fundamentais dos titulares dos dados. Sendo assim as empresas que utilizam de dados pessoais de seus clientes se encontram sujeitas à LGPD e o descumprimento desta lei pode causar prejuízos a empresa, tais como a aplicação de multas pela Autoridade Nacional de Proteção de Dados (ANPD). Portanto, é necessário que as empresas pequenas, médias e grandes busquem a conformidade com a lei. (MÜLLER, 2021).

Empresas e pessoas têm a sua disposição várias tecnologias que possibilitam que diversos recursos de computação sejam fornecidos com alta eficiência e eficácia. Essas tecnologias muito das vezes trabalham com conjuntos de dados muito grandes e complexos para o processamento em *softwares* comuns. Entre essas tecnologias existem aquelas que permitem a descoberta de informações, padrões e correlações de um conjunto de dados em que é possível utilizar técnicas que cruzam as

informações para se chegar ao objetivo, essa área da computação é denominada de ciência de dados. (SILVA, 2019).

Para realizar a ciência de dados é importante a utilização de métodos inteligentes, tais como a mineração de dados. Segundo Castro e Ferrari (2016), a mineração de dados analisa uma base de dados utilizando algoritmos apropriados para alcançar conhecimento que permita tomadas de decisão.

De acordo com Silva (2019), os dados manuseados por meio da análise de um conjunto de dados geralmente possuem informações pessoais e privadas dos titulares dos dados, sendo capaz de implicar em ameaças à privacidade das pessoas. Assim, as empresas que prezam pela sua segurança e de seus clientes devem adotar soluções e estarem em conformidade com leis e regulamentações de proteção de dados como a LGPD. Todos os ativos da empresa devem ser objeto de cuidado da privacidade dos dados das pessoas. Uma das possíveis soluções para segurança da informação e a gestão de riscos quando se trata de dados pessoais é o uso da estratégia de anonimização dos dados. (SILVA, 2019).

A anonimização é composta de técnicas que podem ser utilizadas em um conjunto de dados de modo a evitar a identificação dos titulares. Silva (2019) afirma que a anonimização visa a segurança das informações dos titulares e é utilizada para evitar o vazamento de informações quando existe a necessidade de compartilhar os dados com terceiros. No entanto, a anonimização é realizada antes que os dados sejam compartilhados ou divulgados.

No processo de anonimização dos dados quanto mais estes forem anonimizados menor pode ser a sua utilidade. De acordo com Gabinete de Proteção de Dados Pessoais (GPDP) (2019, p.9-10), ao realizar a anonimização é preciso decidir o grau de compromisso entre utilidade aceitável e tentativa de redução do risco de reidentificação dos dados. Há várias técnicas de anonimização que podem ser utilizadas para o equilíbrio do compromisso de utilidade e privacidade, que neste caso podem ser as técnicas de supressão, generalização, encobrimento de caracteres, agregação e entre outras.

Este trabalho tem como proposta explorar o funcionamento da anonimização dos dados, mineração dos dados e alguns aspectos da LGPD, e assim aplicar as técnicas de anonimização e mineração dos dados em um conjunto de dados, a fim de

anonimizar o conjunto de dados original e para cada etapa da anonimização comparar a acurácia dos conjuntos de dados gerados.

1.2 Justificativa

Os vazamentos de dados relacionados a incidentes de segurança e a ataques *hackers* têm tido destaque nos últimos anos conforme apresentado anteriormente. Por exemplo, no Brasil ocorreram dois grandes vazamentos em 2021, sendo o primeiro deles no mês de janeiro em que a empresa Serasa Experian é a principal acusada pelo vazamento no qual dados pessoais de mais 220 milhões de cidadãos brasileiros foram expostos, incluindo dados de pessoas falecidas. O segundo se deu logo após o primeiro vazamento e expôs dados de 102 milhões de brasileiros e que envolveu duas grandes operadoras do país, Claro e Vivo. (MARI, 2021).

Essas violações de privacidade causam prejuízos aos titulares dos dados e as empresas que tratam os dados. Com base nessas violações vários países do mundo estão criando leis de proteção de dados que obrigam as empresas a adotarem requisitos como forma de proteger e preservar a privacidade das pessoas. Com a utilização de ferramentas que possibilitam a extração de conhecimento em conjuntos de dados muito grandes, a privacidade dos dados se torna mais essencial. A tomada de decisões pode ser feita com base em informações resultantes da análise de dados, proporcionando eficiência e eficácia para as empresas em suas decisões. Com vistas à anonimização dos dados é possível escolher entre privacidade e utilidade dos dados ou o equilíbrio de ambos. No entanto, se a anonimização não for realizada corretamente, com o objetivo de manter a utilidade dos dados, os riscos de reidentificação dos dados pode ser alta. (SILVA, 2019).

Justifica-se estudar este tema pois percebe-se que a anonimização de dados é uma estratégia que pode ser utilizada para preservar e proporcionar a segurança da informação além de ser um grande fator na gestão de riscos e no cumprimento da LGPD. Assim a mineração de dados pode ser uma grande aliada para tomar decisões nos conjuntos de dados anonimizados sem comprometer a privacidade das pessoas.

Diante deste contexto, este estudo pretende responder às seguintes questões de pesquisa:

Q1. A anonimização dos dados contribui ou agrava a qualidade dos resultados dos algoritmos de classificação utilizados no processo de mineração dos dados?

Q2. A anonimização dos dados contribui com segurança da informação, gestão de riscos e cumprimento da LGPD?

Q3. Considerando os dados utilizados nos experimentos, é possível determinar uma simetria entre a privacidade e a utilidade dos dados?

Q4. Com o conjunto de dados anonimizados é possível compartilhar com terceiros sem que haja a reidentificação dos dados?

1.3 Objetivos

1.3.1 Objetivo geral

Este trabalho visa aplicar e analisar as técnicas de anonimização e mineração de dados em um conjunto de dados de crédito. Assim para conjuntos de dados gerados na anonimização, a mineração dos dados é realizada de modo a comparar o efeito desta anonimização nos conjuntos de dados a partir de medidas de acurácia. Como resultados, espera-se evidenciar a eficácia da anonimização e ao mesmo tempo, realizar a mineração nos dados anônimos de forma a obter as informações importantes para a tomada de decisão.

1.3.2 Objetivos específicos

Para atingir o objetivo geral, propõem se os seguintes objetivos específicos:

- Conceituar a Lei Geral de Proteção de Dados;
- Conceituar a anonimização dos dados;
- Conceituar a mineração de dados;
- Aplicar as técnicas de anonimização e mineração de dados em um conjunto de dados.
- Analisar os resultados obtidos no anonimato dos dados a partir da aplicação das técnicas de anonimização utilizando o *software* Amnesia;
- Analisar e comparar os resultados obtidos a partir da aplicação da mineração de dados em cada um dos conjuntos de dados utilizando a ferramenta WEKA;

- Analisar a aplicabilidade da anonimização e mineração dos dados;
- Identificar o impacto da anonimização dos dados no resultado da mineração dos dados.

1.4 Estrutura do trabalho

O presente trabalho apresenta-se estruturado em 5 (cinco) capítulos, sendo o 1 (um) referente a esta introdução. No capítulo 2 (dois) é apresentada toda pesquisa bibliográfica deste trabalho em que são definidos alguns aspectos da Lei Geral de Proteção de dados e conceitos sobre privacidade dos dados, anonimização dos dados e mineração de dados. O capítulo 3 (três) é composto pelos materiais e métodos utilizados neste trabalho. No capítulo 4 (quatro) é descrito os resultados obtidos do experimento com os conjuntos de dados. Por fim o capítulo 5 apresenta as considerações finais.

2 REFERENCIAL TEÓRICO

2.1 Contextualizando a LGPD

A LGPD tornou-se de grande relevância e necessidade. Portanto, busca-se aqui apresentar alguns aspectos da lei e contextualizá-la.

A LGPD (Lei n. 13.709/2018), foi criada com base na lei europeia de proteção de dados, a *General Data Protection Regulation* (GDPR), com objetivo de proteger os dados pessoais de pessoas naturais, ou seja, pessoas físicas. (GARCIA, 2020).

Segundo Peck (2020) no contexto histórico, o motivo que inspirou o surgimento de regulamentações de proteção de dados pessoais está relacionado ao:

> [...] próprio desenvolvimento do modelo de negócios da economia digital, que passou a ter uma dependência muito maior dos fluxos internacionais de bases de dados, especialmente os relacionados às pessoas, viabilizados pelos avanços tecnológicos e pela globalização. (PECK, 2020, p.17).

A LGPD foi sancionada pelo ex-presidente Michel Temer em 14 de agosto de 2018, vigorando no mês de agosto de 2020. Donda (2020) afirma que devido à crise global provocada pela pandemia do novo coronavírus, que praticamente parou todas as operações no Brasil e no mundo as sanções referentes a LGPD foram postergadas. A partir disso foi aprovada uma medida provisória para que as empresas não fossem penalizadas por não se adequarem à lei, devido as recomendações de distanciamento social, prorrogando a aplicação das sanções para agosto de 2021.

De acordo com a Lei nº 13.709, em seu art. 1º a lei se baseia no tratamento de dados pessoais, tratados nos meios físicos ou digitais, por pessoa natural (pessoa física) ou por pessoa jurídica que pode ser de direito público ou privado, com objetivo de proteger os direitos fundamentais de liberdade, privacidade e o livre desenvolvimento da personalidade das pessoas.

Os fundamentos da disciplina de proteção de dados pessoais são descritos no art. 2º da Lei nº 13.709, e contém sete (7) incisos, que devem ser cumpridos a fim de proporcionar transparência, segurança, responsabilidade, dedicação e

desenvolvimento em proteção de dados pessoais. A Figura 1 mostra os pilares estabelecidos quando se cumprem os fundamentos da LGPD.

Figura 1: Fundamentos da LGPD

Fonte: OBJETIVO (2018).

No art. 6ª da Lei nº 13.709, é determinado que as atividades relacionadas ao tratamento dos dados pessoais devem observar a boa-fé e os dez princípios propostos a serem cumpridos, sendo eles:

- Finalidade: propósito legitimo da coleta e dos tratamentos de dados informados ao titular.
- Adequação: o tratamento deve ser compatível com a finalidade.
- Necessidade: limitar o tratamento ao mínimo necessário.
- Livre acesso: garantir ao titular consulta (gratuita), duração e integralidade dos dados.
- Qualidade dos dados: exatidão, clareza e relevância dos dados de acordo com a necessidade e para garantir a finalidade.
- Transparência: garantir aos titulares informações claras, precisas e facilmente acessíveis sobre realização do tratamento e os respectivos agentes de tratamento;
- Segurança: adotar medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;

- Prevenção: adotar medidas de prevenir a ocorrência de danos;
- Não discriminação: não permitir a realização do tratamento para fins discriminatórios ilícitos ou abusivos;
- Responsabilização e prestação de contas: demostrar a adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

Portanto, diante deste contexto percebe-se a importância do cumprimento da LGPD por parte das empresas que tratam dados pessoais e que estão na busca pela segurança e responsabilidade com os dados das pessoas.

Logo, após analisar alguns aspectos da LGPD é relevante compreender o que são os dados pessoais e como a LGPD determina e classifica os requisitos para realizar o tratamento dos dados pessoais.

2.1.1 Tratamento dos dados pessoais

A Lei aplica-se a qualquer operação de tratamento de dados pessoais, independentemente do meio, do país de sua sede ou do país onde estejam localizados dos dados, sendo uma lei que se situa fora dos limites territoriais do país. (DONDA, 2020).

Antes de compreender o que são dados pessoais é preciso entender quem é o portador destes dados. “O titular dos dados pessoais é toda pessoa natural a quem se referem os dados pessoais que são objeto de tratamento.” (SOARES, 2019, p. 7).

Donda (2020), apresenta a definição dos tipos de dados presentes no artigo 5º da LGPD. Sendo eles:

- Dado pessoal: qualquer informação relacionada a um indivíduo, uma pessoa identificada ou identificável, tais como nome, CPF, endereço, etc.
- Dado pessoal sensível: se refere a origem racial ou étnica, convicção religiosa, opinião política, opção sexual, dado genético, dentre outros.
- Dado anonimizado: é qualquer dado pertencente a um indivíduo que não pode ser identificado, o que acontece no momento do tratamento, podendo-se utilizar de mecanismos para alterar as informações, de modo que esses dados perdem a associação ao indivíduo.

No art. 5º da lei nº 13.709, fica estabelecido que o tratamento de dados é toda operação realizada com dados pessoais, como as que se referem a coleta, utilização, produção, distribuição, acesso, reprodução, transmissão, processamento, arquivamento, armazenamento, eliminação, modificação.

De acordo com Maciel (2019), o titular dos dados é quem define como seus dados pessoais devem ser tratados, caso a base legal utilizada seja o consentimento. No entanto, o consentimento é apenas uma das bases legais para validar o tratamento dos dados pessoais. A Lei nº 13.709, determina dez requisitos legais para realizar o tratamento de dados pessoais de acordo com o art. 7º as bases legais para realizar o tratamento dos dados são:

- O titular dos dados deve consentir com o tratamento dos dados;
- Os dados são coletados para cumprir com obrigações legais ou regulatórias por parte do controlador;
- Pela administração pública para fins de administração pública, defesa nacional, investigações ou segurança do estado;
- Com base em estudos realizados por órgãos de pesquisa, porém que sempre que possível os dados sejam anonimizados;
- Quando necessário para execução de contrato ou de procedimentos preliminares relacionados a contrato, na qual o titular faz parte do contrato e não pode cancelar até que dure o contrato;
- Para exercício regular de direitos em processo judicial, administrativo ou arbitral;
- Para proteção da vida ou da incolumidade física do titular ou terceiro;
- Para tutela da saúde, desde que os procedimentos sejam realizados por profissionais de saúde, serviços ou autoridade sanitária;
- Quando necessário para atender os interesses legítimo do controlador ou de terceiros;
- Para proteção do crédito.

E ainda é preciso levar em consideração que o tratamento dos dados pessoais deve ser interrompido nas seguintes hipóteses de acordo com art.15º da LGPD, tal como exposto por Maciel (2019):

- Quando a finalidade foi alcançada ou que os dados deixaram de ser necessários ou pertinentes ao alcance da finalidade específica desejada;
- Fim do período de tratamento;
- Por comunicação do titular, inclusive no exercício de seu direito de revogação do consentimento; ou
- Determinação da autoridade nacional, quando houver violação ao disposto nesta Lei.

Porto (2020) ressalta um importante critério, que em caso de desconformidade com a LGPD, atribuída ao tratamento dos dados pessoais, os agentes de tratamento estarão sujeitos às penalidades administrativas dispostas no art.52 da lei que somente será aplicada pela Autoridade Nacional de Proteção de Dados (ANPD), na qual a infração pode-se apresentar defesa.

Desta forma, compreender a relevância dos dados pessoais e a forma como estes podem influenciar a gestão de todos os processos dentro das empresas se faz necessário pois, pensando de forma sistêmica todos os atributos da LGPD devem estar em conformidade a fim de estabelecer uma cultura de segurança e conformidade satisfatória.

Outro aspecto relevante ao se falar sobre a estruturação e implementação da LGPD são os atores da mesma, ou seja, os envolvidos no tratamento dos dados pessoais.

2.1.2 Atores da LGPD

Os atores da LGPD são os principais envolvidos no tratamento de dados pessoais. Portanto, busca-se aqui apresentar suas responsabilidades e atribuições definidas na lei.

De acordo com a Lei nº 13.709, art. 5º incisos V ao VIII é apresentado os principais indivíduos envolvidos no tratamento dos dados, sendo estes: O titular dos dados, controlador, operador e o encarregado dos dados. “O titular dos dados

pessoais e pessoais sensíveis é o indivíduo que a lei visa proteger e é identificado como portador dos dados pessoais que são objeto de tratamento.” (PALUDETTO; BARBIERI, 2019, p.3).

Em aspectos de hierarquia o controlador é quem toma as decisões, assim o operador é aquele que obedece às decisões tomadas. Logo, “o controlador pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais” (BRASIL, 2018). É obrigação do controlador elaborar o Relatório de Impacto a Proteção de Dados Pessoais (RIPD).

> Relatório de impacto à proteção de dados pessoais: documentação do controlador que contém a descrição dos processos de tratamento de dados pessoais (Ciclo dos Dados) que podem gerar riscos (Risk Assessment) às liberdades civis e aos direitos fundamentais, bem como medidas, salvaguardas e mecanismos de mitigação de risco, tais como mapeamentos, treinamentos, auditorias, alterações de contrato e criação de políticas de proteção de dados. (SOARES, 2019, p.9).

De acordo com Maciel (2019) as obrigações do operador devem estar definidas de acordo com as decisões do controlador, a fim de atender as suas necessidades com agilidade e comprometimento, no intuito de evitar punições ao controlador. Assim “o operador pessoa natural ou jurídica, de direito público ou privado, que realiza o tratamento de dados pessoais em nome do controlador” (BRASIL, 2018).

Segundo Porto (2020) para desempenhar as funções de encarregado dados, o profissional deve ter conhecimentos de privacidade e proteção de dados, além da legislação de proteção de dados para adotar as devidas providências decorrentes de comunicações com a ANPD, e ainda orientar os colaboradores sobre as medidas a serem tomadas referentes a proteção de dados. Com isso “o encarregado de dados pessoa indicada pelo controlador e operador para atuar como canal de comunicação entre o controlador, os titulares dos dados e a ANPD.” (BRASIL,2018).

Os agentes de tratamento são o controlador e o operador. No dizer de Donda (2020), são responsáveis pela segurança e privacidade dos dados, além de serem os atores responsáveis por indicar o encarregado dos dados. O Quadro 1 apresenta um comparativo de obrigações gerais dos agentes de tratamento em que o controlador e operador devem seguir para estarem em conformidade com a lei:

Quadro 1: Comparativo de obrigações gerais

	OBRIGAÇÕES GERAIS	
	CONTROLADOR	OPERADOR
Limites para tratamento	Tratar dados com base legal definida	Tratar dados conforme e propósitos definidos pelo controlador
Registros	Registro das atividades	Registro das atividades
Relatório de Impacto	Elaborar relatório de impacto com boa prática e força legal	Elaborar relatório de impacto com boa prática
Encarregado	Indicar encarregado	Definir em contrato pessoa responsável pela comunicação com o controlador
Direitos dos titulares	Atender aos direitos dos titulares	Colaborar com o controlador
Incidentes	Comunicar à autoridade nacional e ao titular a ocorrência de incidente de segurança que possa acarretar risco ou dano relevante aos titulares.	Informar o controlador casos de incidentes
Boas práticas de segurança	Adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situação acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de	Adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situação acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de

	tratamento inadequado ou ilícito.	tratamento inadequado ou ilícito.
Programa de Governança em privacidade	Implementar programa de Governança em Privacidade, observadas a estrutura, a escala e o volume de suas operações bem como a sensibilidade dos dados tratados e a probabilidade e a gravidade dos danos para os titulares.	Receber e estar ciente do Programa de Governança adotado pelo controlador.

Fonte: Adaptado de Maciel (2019, p.1329)

Desta forma, pode-se observar a inclusão de uma cultura interna e externa, apresentando-se por meio destes indivíduos (atores) estabelecidos pela LGPD para realizar o tratamento dos dados pessoais. Por exemplo, o titular dos dados é um indivíduo externo da organização, que possui seus princípios e costumes, que podem interferir na cultura de segurança da informação da empresa.

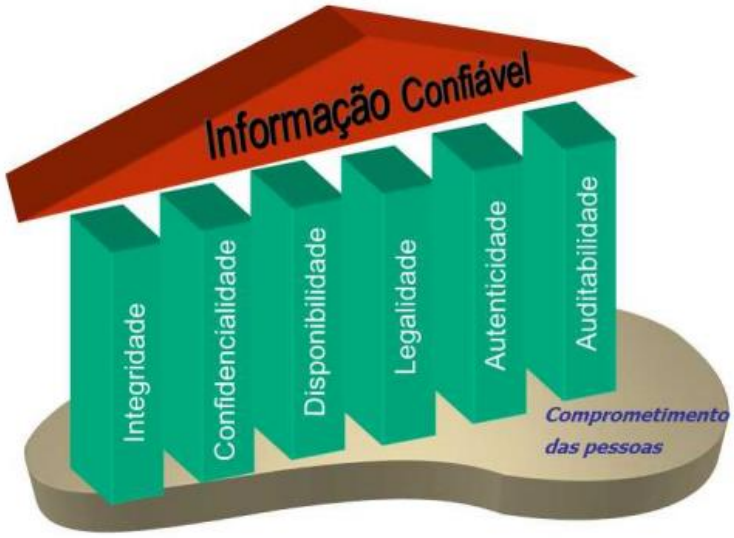
Portanto, é relevante analisar todos os aspectos que envolvem estes sujeitos sociais pois são de extrema importância. A partir dos ativos é importante estabelecer a segurança da informação afim de proteger tais informações de acessos não autorizados.

2.1.3 Segurança da informação

Um dos aspectos que se deve levar em consideração ao mencionar LGPD é a segurança da informação, pois quando se trata de proteção de dados, toda a estrutura de segurança deve ser pensada e avaliada.

De acordo com Fontes (2020) a segurança da informação tem como objetivo garantir uma informação confiável, através da integridade, confidencialidade, disponibilidade, legalidade, autenticidade e auditabilidade. A Figura 2 apresenta estes objetivos, tendo como base o comprometimento das pessoas.

Figura 2: Os seis pilares da segurança da informação.



Fonte: FONTES (2020).

Na LGPD, é determinado como obrigação que “Os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de tratamento inadequado ou ilícito” (BRASIL, 2018).

Segundo Maciel (2019), a maioria dos incidentes relacionados à segurança estão interligados a falhas humanas.

> [...] de nada adianta o investimento em softwares avançadíssimos se os colaboradores não estão envolvidos com a proteção à privacidade. Não à toa que um programa de governança em privacidade é sempre construído a partir do topo, conscientizando a alta gestão, para depois ir conscientizando toda a equipe, para que o que for previsto no programa seja de fato incorporado às práticas diárias. (MACIEL, 2019, p. 1421).

Os colaboradores precisam ser treinados em segurança da informação. Para isto é necessário que cada um destes conheçam a Política de Segurança da Informação (PSI), estabelecida pela empresa. (FONTES, 2020).

De acordo com Fontes (2020), a segurança da informação deve ser realizada por um profissional capacitado que tenha experiência em gestão da segurança, com uma visão geral das ameaças e riscos que cercam as informações. Inicialmente uma

gestão e governança pode ser estruturada com base em Normas Brasileiras (NBR) da família ISO/IEC (Organização Interacional de Padronização/International Electrotechnical Commission) 27000 que são dimensionadas a segurança da informação.

Soares (2019) afirma que o termo *Privacy By Desing* (Privacidade desde a concepção) é um princípio de governança previsto na LGPD, que estabelece que as empresas devem integrar a privacidade a todas as etapas de um determinado projeto, sistema ou negócio. Tal afirmação corrobora com a ideia de que “as medidas técnicas e administrativas devem ser pensadas desde a concepção do produto ou do serviço até a sua execução.” (MACIEL, 2019, p. 1451).

Donda (2020) afirma que para a gestão da segurança da informação é necessário estabelecer e implementar a segurança da informação com bases nos seis pilares apresentados anteriormente. Logo faz parte da gestão de segurança da informação:

- Criar uma política de segurança da informação;
- Coordenar as atividades de segurança da informação;
- Fazer a gestão de ativos;
- Proteger e classificar a informação;
- Garantir a segurança lógica e física do ambiente;
- Acompanhar a gestão de mudanças;
- Gerenciar a segurança e o controle de acesso;
- Detectar atividades não autorizadas por meio do monitoramento ambiente;
- Fazer as análises de vulnerabilidades;
- Fazer a gestão de incidentes de segurança;
- Implementar um plano de continuidade do negócio;
- Manter conformidade com normas e leis.

Desta forma, “a política de segurança da informação é um documento que tem como função orientar e estabelecer diretrizes sobre a proteção da informação.” (DONDA, 2020, p. 31)

Portanto, estabelecer e compreender a segurança da informação é indispensável para todas as empresas. A partir do momento em que se integra a

segurança da informação aos colaboradores de uma empresa, observa-se inúmeros benefícios. Espera-se que todos estejam envolvidos na conformidade dos pilares da segurança da informação, pois quando se tem tais atributos internamente (integridade, confidencialidade, legalidade, autenticidade, auditabilidade e disponibilidade), a empresa passa uma nova imagem que agrega valor e comprometimento em relação a segurança. Desta forma a segurança da informação irá influenciar todos os âmbitos organizacionais e passa a ser levada em consideração dentro de todos os processos.

Além da segurança da informação e de todas as variáveis envolvidas quando se fala em sua gestão, deve-se considerar as possiblidades quando algo não sai como planejado. Deste modo, faz-se necessário compreender e estabelecer a gestão de risco e todo o contexto envolvido.

2.1.4 Gestão de riscos

A partir de uma segurança da informação implementada, precisa-se observar, gerenciar, identificar ou minimizar os possíveis riscos que podem ocorrer com os ativos da empresa que se encontram vulneráveis. Estes aspectos são essenciais na gestão de riscos e cumprimento da LGPD. Busca-se aqui compreender tal necessidade e exemplificá-la.

Donda (2020) apresenta de forma simplificada as informações presentes no nos temas de Segurança e do Sigilo de Dados, Das Boas Práticas e da Governança, estabelecidos na lei, na qual expõe a proteção de dados pessoais e evidência a necessidade de controle e monitoramento pelos agentes de tratamento, além de utilizar de mecanismos para mitigar os riscos.

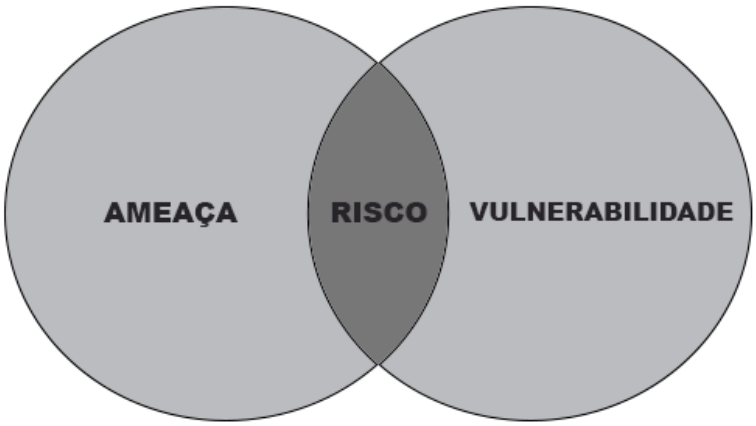
> Art. 50. Os controladores e operadores, no âmbito de suas competências, pelo tratamento de dados pessoais, individualmente ou por meio de associações, poderão formular regras de boas práticas e de governança que estabeleçam as condições de organização, o regime de funcionamento, os procedimentos, incluindo reclamações e petições de titulares, as normas de segurança, os padrões técnicos, as obrigações específicas para os diversos envolvidos no tratamento, as ações educativas, os mecanismos internos de supervisão e de mitigação de riscos e outros aspectos relacionados ao tratamento de dados pessoais. (BRASIL, 2018).

De acordo com Lopes (2020), os riscos sempre se apresentam presente nos negócios com intuito de arruinar os negócios, seja por meio de impactos financeiros,

vazamento de dados, quebra de sigilo e entre outros. Assim a implementação de um sistema de gestão de riscos, tem a finalidade de promover uma análise de riscos e medidas de prevenção com base em regras, condutas e ética aplicáveis a tecnologia da informação.

Segundo Donda (2020), é preciso conhecer os riscos e a quais ativos da organização eles se encontram associados. Desta forma a “avaliação e análise dos riscos são muito importantes para dar visibilidade à situação dos ativos, assim como priorizar os investimentos e proteger os ativos da melhor maneira” (DONDA, 2020, p. 68). A Figura 3 apresenta a existência do risco a partir dos aspectos de ameaça e vulnerabilidade encontrado nos ativos.

Figura 3: Ameaças e vulnerabilidades associadas geram riscos



Fonte: Adaptado de Donda (2020, p. 68).

O risco não existe sem as ameaças e vulnerabilidades associadas. Para conhecer um risco é preciso entender o que está envolvido a ele e se é um ativo organizacional ou não. Sendo assim uma ameaça é qualquer condição que possa causar dano ou perda e a vulnerabilidade é uma fraqueza explorável, como falhas no desenvolvimento de um software ou qualquer configuração incorreta ou mal executada. (DONDA, 2020).

De acordo com Brasil (2020), os riscos identificados podem ser classificados de acordo com sua probabilidade de ocorrência e com base no possível impacto. A Tabela 1, apresenta parâmetros escalares para se classificar os riscos.

Tabela 1: Parâmetros Escalares

CLASSIFICAÇÃO	VALOR
Baixo	5
Moderado	10
Alto	15

Fonte: BRASIL (2020).

Os parâmetros escalares podem ser utilizados para identificar os riscos. Os valores da classificação (baixo, moderado e alto) são definidos pelo encarregado de realizar a gestão dos riscos. Assim os parâmetros escalares são empregues para representar os níveis de probabilidade e impacto que, ao ser multiplicado tem como resultado o nível de risco que auxiliará na aplicação de medidas de segurança. (BRASIL, 2020).

Assim, a partir desta classificação é possível definir o nível dos riscos, para que se possa adotar medidas. É “importante reforçar que as medidas para tratar os riscos podem ser: de segurança, técnicas ou administrativas.” (BRASIL, 2020, p. 41)

Nas palavras de Donda (2020), precisa-se conhecer o ambiente para poder proteger, no sentido de que no decorrer do mapeamento dos dados seja possível definir e identificar os riscos que rodeiam os ativos da organização. A Figura 4 apresenta alguns exemplos de ativos organizacionais.

Figura 4: Ativos de uma empresa

Fonte: BRASIL (2020).

A “[...] gestão de risco informático passa sempre pela chamada análise de riscos, que visa a apurar, no caso concreto, a quais riscos o ambiente tecnológico está exposto e quais são as medidas necessárias para o seu controle.” (BIONI et al., 2020, p.359)

Portanto, a gestão de riscos é uma das etapas imprescindíveis quando se busca ter um controle dos riscos e sua implementação tende a contribuir no fortalecimento da segurança da informação de forma adequada. Entretanto, é preciso estar atento a todos os elementos envolvidos na gestão, caso contrário uma análise mal estruturada pode levar a empresa a lidar com consequências negativas.

Compreender a importância da gestão de riscos vai muito além de meras análises, é preciso conscientizar todos os colaboradores sobre a importância de todo este planejamento relacionado a proteção dos dados.

Portanto, quando se trata de dados pessoais e riscos envolvendo seus atributos, a anonimização de dados se torna uma estratégia forte na gestão de riscos.

2.2 Anonimização dos dados - conceitos

A anonimização de dados é de suma importância para alcançar a privacidade dos dados e a segurança da informação. Desta forma compreender o conceito de anonimização de dados é o primeiro passo para posteriormente entender as técnicas.

As empresas recebem dados pessoais constantemente para atender as suas necessidades e de seus clientes. Com a vigência da LGPD, o compartilhamento dos dados deve atender aos requisitos da lei além de proteger a privacidade das pessoas. Entretanto, as empresas que compartilham dados devem considerar a anonimização e estar atentas para que os dados não sejam reidentificados. Neste sentido, entender como funciona a anonimização de dados é vital. (K-ANONIMATO, 2017).

De acordo com Silva (2019), a anonimização dos dados é uma das principais estratégias para obter a busca pela privacidade, desde que aplicada corretamente. Os procedimentos e modelos de anonimização contribuem na preservação da identidade do indivíduo. O dado anonimizado é um dado pessoal que pertence a um indivíduo, que não pode ser reidentificado após à aplicação de técnicas para anonimizar os dados originais.

Além disso, Carloto (2020) complementa ao dizer que na LGPD a anonimização dos dados não é exigida. Porém, pode-se utilizar de técnicas e modelos de anonimização para evitar a identificação do indivíduo titular dos dados. Assim as empresas que tratam dados anonimizados se resguardam das sanções e normas aplicadas por legislações de proteção de dados.

Silva (2019) afirma ainda que no processo de anonimização é importante definir o conjunto de dados que deve ser anonimizado e definir os procedimentos que serão aplicados nos dados. Se os dados forem divulgados ou compartilhados com terceiros, é necessário que estes estejam classificados de acordo com a sensibilidade da informação.

De acordo com Camenisch; Fischer-Hübner; Rannenberg (2011 apud SILVA, 2019, p.28) a classificação dos dados é dividida em três atributos, sendo eles:

- Atributos identificadores: nome, CPF, Registro Geral (RG);
- Atributos semi-identificadores: data de nascimento, Código de Endereçamento Postal (CEP), cargo, tipo sanguíneo;
- Atributos sensíveis: salário, religião, sexo.

Em virtude dos aspectos abordados, é relevante assimilar a anonimização dos dados e sua contribuição na busca pela privacidade dos dados, pois as empresas que realizam a anonimização dos dados aperfeiçoam a segurança da informação e geram mais confiança com clientes e parceiros. Logo, é necessário apresentar as técnicas de anonimização utilizadas na busca pela privacidade.

2.2.1 Técnicas de anonimização

As técnicas de anonimização são essenciais para cumprir a meta de anonimização dos dados e proteger a privacidade das pessoas. Há diferentes tipos de técnicas de anonimização de dados, e cada uma pode ser utilizada para modificar os dados de acordo com a necessidade. As principais técnicas de anonimização utilizadas são: supressão, generalização, encobrimento de caracteres e agregação. Cada uma delas tem suas especificidades, como é mostrado a seguir.

2.2.1.1 Supressão

A técnica de Supressão refere-se à remoção de uma ou várias colunas de dados no conjunto de dados. Essa técnica é utilizada quando um atributo não é necessário no conjunto de dados anonimizados, ou quando o dado não pode ser anonimizado utilizando de outras técnicas. Um conjunto de dados é estruturado em uma tabela que possui linhas e colunas em que cada linha representa um conjunto de dados e as colunas representam os atributos de cada tipo de dado. Na supressão a coluna deve ser eliminada, não podendo apenas ocultar a coluna. (GPDP, 2019).

A supressão normalmente é aplicada nos dados que identificam o titular dos dados. A Tabela 2 mostra um conjunto de dados que contém dados pessoais que identificam atletas e seus respectivos esportes, sexo e vitórias em competições. Com a aplicação da supressão os atributos CPF e Nome são excluídos, gerando um novo conjunto de dados conforme apresentado na Tabela 3.

Tabela 2: Tabela de dados originais

CPF	NOME	Sexo	Esporte	Vitórias em Competições
12345678910	Alef S Duarte	Masc	Futebol	12
10111213141	Anderson M Santos	Masc	Natação	5
45124556890	Fernando S Leite	Masc	Futebol	12
78789855220	Eduardo Miranda	Masc	Ginástica	5

Fonte: Elaborado pelo autor.

Tabela 3: Tabela de dados com aplicação da técnica de supressão

Sexo	Esporte	Vitórias em Competições
Masc	Futebol	12
Masc	Natação	5
Masc	Futebol	12
Masc	Ginástica	5

Fonte: Elaborado pelo autor.

Observe a Tabela 3, que a aplicação da supressão contribui com o anonimato dos dados não sendo possível identificar os titulares dos dados. No entanto a supressão dos dados é uma das técnicas de anonimização que pode ser utilizada para

alcançar o anonimato dos dados e sua aplicação requer cautela pois se trata de uma técnica de forte aplicação em que não é possível recuperar um atributo removido.

2.2.1.2 Encobrimento de caracteres

A técnica de encobrimento de caracteres tem como objetivo alterar os caracteres parcialmente, ou seja, ocultar uma parte da cadeia de caracteres com intuito de que o grau de anonimização seja aumentado e assim impossibilitar a recuperação dos dados que foram encobertos. (GPDP, 2019).

Para exemplificar a técnica de encobrimento de caracteres é utilizado a estrutura do número de matrícula de discentes da Pontifícia Universidade Católica de Goiás, que é um conjunto numérico constituído de catorze algarismos que identificam o aluno dentro da universidade. A Figura 5 mostra como o número de matrícula está estruturado para os alunos da universidade. Os quatro primeiros algarismos se referem ao ano de ingresso do estudante, o quinto algarismo ao semestre de ingresso, do sexto até o nono algarismo é associado ao número do curso, do decimo ao decimo terceiro algarismo é definido o número que identifica o aluno dentro do curso e o último número é o digito de controle. (MANUAL, 2021).

Figura 5: Estrutura do número de matrícula

Fonte: Manual do aluno (MANUAL, 2021)

Com base no número de matrícula, o encobrimento de caracteres pode ser utilizado a fim de alcançar a privacidade do aluno dentro da universidade. Quando aplicado a técnica de anonimização o número de matrícula 2018.1.0001.3009-2 passa a ser 2018.1.0001.****-*. Desta forma, mantém-se a privacidade e a não identificação

do aluno dentro da instituição de ensino e assim as notas em disciplinas podem ser compartilhadas com terceiros, desde que não permita a reidentificação do aluno.

A Tabela 4 apresenta as notas de alunos que tiveram suas matrículas com os caracteres encobertos, com o objetivo de evitar a identificação dos alunos.

Tabela 4: Técnica de encobrimento de caracteres aplicada no atributo matrícula

MATRÍCULA	P1	P2	N1	P3	P4	N2	MÉDIA
2013.1.0028.****-*	1,3	2,0	0,7	2,9	2,7	1,7	2,4
2014.1.0033. ****-*	4,1	6,2	2,1	5	8	3,9	6
2015.2.0028. ****-*	8	7.5	3,1	6.2	5	3,4	6,5
2016.1.0028. ****-*	5	6	2,2	7.1	6.2	4	6,2
2017.2.0033. ****-*	4	9,5	2,7	3.4	6.9	3,1	5,8
2018.1.0028. ****-*	9.1	9	3,6	5.7	8	4,1	7,7
2018.2.0033. ****-*	5,5	5	2,1	2.9	7	3	5,1

Fonte: Elaborado pelo autor.

2.2.1.3 Generalização

A técnica generalização reduz a precisão dos dados, e é aplicável em valores que podem ser generalizados e úteis para alcançar o objetivo pretendido. Na generalização é possível conceber faixas etárias em dimensões apropriadas para as quais o valor original deve estar entre a faixa estabelecida. (GPDP, 2019).

Por exemplo, ao generalizar um conjunto de dados que contém o atributo idade é possível definir a faixa numérica para acomodar a idade. Observando a Tabela 5, se tem o atributo idade generalizado, não sendo possível identificar a idade daquele indivíduo.

Tabela 5: Atributo idade generalizado

ID	NOME	IDADE
1	3566	21-30
2	3515	31-40
3	2544	41-50

4	7855	51-60

Fonte: Elaborado pelo autor.

2.2.1.4 Agregação

Sua função é converter um conjunto de dados de uma lista para valores resumidos e proteger os dados dos indivíduos de reidentificação. No processo de agregação, consultas realizadas não podem trazer registros exclusivos que identificam o indivíduo. (SILVA, 2019).

Um exemplo de aplicação da técnica de agregação é apresentado na Tabela 7, na qual a técnica é aplicada na Tabela 6 que ilustra um conjunto de dados pessoais, sem nenhum tratamento. Após a aplicação da agregação uma nova tabela de dados é gerada conforme a Tabela 7.

Tabela 6: Tabela de dados pessoais original

Nome	**Salário mensal ($)**	**Carros**	**Motos**
Glenda Barbosa	5000	1	1
João Batista	2000	1	1
Luís Otávio	1000	0	1
Felipe Santos	1000	1	1
Matheus Silva	10000	2	3
Vicente Júnior	5000	1	2
Maria José	5000	1	0
Antônio Silva	2000	1	0
Paulo João	10000	2	3
Francisca Ferreira	1000	0	0
Josefa Talles	2000	1	1
Rosa Duarte	5000	1	2

Fonte: Elaborado pelo autor.

Tabela 7: Conjunto de dados anonimizados utilizando da técnica de agregação

Salário mensal ($)	Soma da quantia de carros	Soma da quantia de motos	Quantidade de pessoas com estes requisitos
500-1500	1	2	3
1500-2500	3	2	3
4500-5500	4	5	4
9500-10500	4	6	2

Fonte: Elaborado pelo autor.

É importante destacar que existem outras técnicas de anonimização que podem ser utilizadas, mas foram discutidas apenas as principais técnicas a fim de exemplificar suas aplicações em um conjunto de dados. Assim, a partir das técnicas, os modelos de anonimização são descritos.

2.3 Modelos de Anonimização

A partir das técnicas de anonimização, os modelos de anonimização são definidos e estão disponíveis para contribuírem com a privacidade dos dados. Nesta seção são apresentados três principais modelos de anonimização e suas características.

2.3.1 O modelo k–anonimato

O k-anonimato é um modelo de privacidade utilizado para evitar a reidentificação de dados anônimos em um conjunto de dados, e para que esta condição seja alcançada é necessário que haja pelo menos k-indivíduos no conjunto de dados e que os seus atributos sejam compartilhados no intuito de que tornem-se identificadores para outros indivíduos. (K-ANONIMATO, 2017).

Com o k-anonimato, informações em um conjunto de dados original podem ser anonimizadas para que um intruso ou terceiros não consigam determinar a identidade do titular dos dados. As técnicas existentes no k-anonimato são de generalização, supressão e entre outras. As probabilidades de que um indivíduo seja identificado em

um conjunto de k-anônimos é de 1 dividido por k. Valores altos de k implicam em uma probabilidade menor de reidentificação. Porém, os dados são distorcidos, pois haverá maior perda de informações devido ao k-anonimato. (EMAM; DANKAR, 2008).

De acordo com Eman e Dankar (2008), a preocupação do k-anonimato é que um único individuo seja reidentificado em um conjunto de dados anônimo. Para que isso aconteça tem-se dois cenários de reidentificação, que são:

- Reidentificar um indivíduo específico: O invasor ou detentor dos dados sabe que um determinado indivíduo existe no conjunto de dados anônimos e quer descobrir a qual tupla de dados pertence ao indivíduo.
- Reidentificar um indivíduo arbitrário: O invasor ou detentor dos dados, não se importa com qual individuo está sendo reidentificado, porém está interessado na reidentificação de qualquer indivíduo, a fim de mostrar para a organização que realiza o tratamento dos dados, que os dados podem ser reidentificados.

A Figura 6, apresenta dois conjuntos de dados no qual o banco de dados original (*Original Database Not Disclose*) é anonimizado por k-anonimato utilizando das técnicas de supressão e generalização.

Figura 6: Tabela de dados anonimizados por k-anonimato

Original Database Not Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	ve+
2	Alan Smith	Male	1962	ve-
3	Alice Brown	Female	1955	ve-
4	Hercules Green	Male	1959	ve-
5	Alicia Freds	Female	1942	ve-
6	Gill Stringer	Female	1975	ve-
7	Marie Kirk	Female	1966	ve+
8	Leslie Hall	Female	1987	ve-
9	Bill Nash	Male	1975	ve-
10	Albert Blackwell	Male	1978	ve-
11	Reverly Mc	Female	1964	ve-
12	Douglas Henry	Male	1959	ve+
13	Freda Shields	Female	1975	ve-
14	Fred Thompson	Male	1967	ve-

2- Anonymization

	QUASI-IDENTIFIERS		
ID	Gender	ear of Birt	Test Result
1	Male	1950-1959	ve+
2	Male	1960-1969	ve-
4	Male	1950-1959	ve-
6	Female	1970-1979	ve-
7	Female	1960-1969	ve+
9	Male	1970-1979	ve-
10	Male	1970-1979	ve-
11	Female	1960-1969	ve-
12	Male	1950-1959	ve+
13	Female	1970-1979	ve-
14	Male	1960-1969	ve-

Disclosed (k-Anonymized) Database

Fonte: Elaborado pelo autor com base em EMAM; DANKAR (2008).

Como resultado da anonimização tem-se a tabela (*Disclosed (k-Anonymized) Database*), na qual é possível observar que após a aplicação da técnica de supressão no atributo identificador *Name* e generalização no atributo semi-identificador *Year of Birth* não é possível reidentificar os titulares dos dados ao realizar uma busca na tabela anonimizada. Algumas tuplas com informações foram removidas devido ao k-anonimato, como forma de evitar a reidentificação dos dados.

2.3.2 O modelo l–diversidade

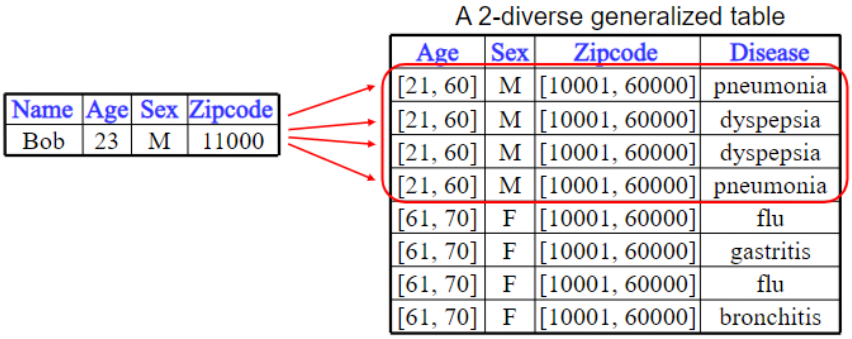
O modelo l-diversidade é uma extensão do k-anonimato com intuito de garantir que ataques determinísticos de inferência não ocorram. No entanto, seu objetivo é

limitar as classes de equivalência com fraca variabilidade dos atributos. Portanto, quando os dados se encontram bem distribuídos, a l-diversidade pode ser utilizada para proteger os dados contra-ataques de inferência, mas se os atributos forem distribuídos no conjunto de dados de forma desigual ou pertencer a uma quantidade reduzida de valores ou de significados semânticos esta técnica pode estar sujeita a ataques de inferência. (GRUPO DE TRABALHO DO ARTIGO 29 (GT), 2014).

De acordo com Silva (2019), o modelo l-diversidade implica l-anonimato, sendo similar ao modelo k-anonimato, pois os dados precisam ser anonimizados para logo após serem l-diversificados. Contudo, um atributo sensível não pode possuir granularidade de valores intragrupo maior que 1 dividido por l.

A Figura 7 apresenta uma tabela com conjunto de dados anonimizados utilizando a técnica l-diversidade. Pode-se observar que o indivíduo Bob não pode ser identificado com base nos atributos *Age*, *Zipcode* e *Diasease*.

Figura 7: Tabela anonimizada por l-diversidade



Fonte: SUNDARESAN (2016).

Com base na Figura 7 é possível observar que o indivíduo procurado (Bob) se enquadra na faixa generalizada pelo k-anonimato. Observando o atributo Disease, não é possível deduzir qual o diagnóstico (pneumonia ou dispepsia) do indivíduo, pois a probabilidade é 1 dividido por 2 no mesmo grupo de equivalentes. (SILVA, 2019).

2.3.3 O modelo t–proximidade

Para Brito e Machado (2017), o modelo t-proximidade é um aperfeiçoamento da l-diversidade, que visa assegurar que a distribuição dos dados com atributo

sensível em cada classe de equivalência esteja próxima a distribuição global. A distância máxima entre as classes e a distribuição global é definida pelo parâmetro t.

Silva (2019) afirma que o k-anonimato define como a t-proximidade requisita que a distribuição de um atributo sensível em qualquer classe de equivalência seja próxima da distribuição do atributo na tabela global, conforme exposto na Tabela 8.

Tabela 8: Tabela de dados anonimizada por t-proximidade

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Fonte: LI (2006).

Com base na análise de Silva (2019, p.34), “a distribuição dos atributos sensíveis (*Disease*) dentro de cada grupo de semi-identificadores deve estar “próxima” de sua distribuição em todo banco de dados original”.

A partir dos modelos de anonimização evidenciados, é importante destacar que com a junção das técnicas aplicadas, a anonimização dos dados pode ser ampliada. Porém, o conhecimento das técnicas é essencial para conseguir alcançar o anonimato. Assim os ataques deixam de ser possíveis em uma base de dados anonimizados. Portanto, é relevante apresentar alguns *softwares* utilizados para anonimização, com objetivo de aplicar as técnicas dos modelos de anonimização em conjunto de dados.

2.3.4 *Softwares* para anonimização

A partir das técnicas e modelos de anonimização, diversos *softwares* podem auxiliar no processo de anonimização e suportar uma ampla variedade de métodos e modelos para aplicação nos dados. Nesta seção, busca-se descrever duas ferramentas *Open Source*, utilizadas para anonimização de dados. Para realizar a

descrição das características destas ferramentas são considerados os recursos disponibilizados de acordo com suas documentações.

A ferramenta ARX é um *software* de código aberto utilizado para o anonimato de dados pessoais e que é capaz de lidar com grandes conjuntos de dados. O ARX utiliza de um algoritmo de pesquisa globalmente otimizado e bastante eficiente para que os dados sejam transformados com a generalização e supressão. (PRASSER et al., 2020).

De acordo com Prasser e Kohlmayer (2015), o ARX é apresentado como uma ferramenta de anonimização que suporta métodos de controle de divulgação estatísticos, fornecendo:

- Modelos para analisar riscos de reidentificação;
- Anonimização baseada em risco;
- Utiliza de modelos de anonimização como: k -anonimato, l-diversidade, t-proximidade e entre outros;
- Métodos manuais e automatizados para utilidade dos dados;
- Utiliza das técnicas de generalização, supressão e agregação.

Segundo Silva (2019), o *software* de anonimização ARX possui uma grande quantidade de recursos disponíveis para uso:

> A ferramenta ARX é capaz de anonimizar dados em big data, pois suporta o uso de milhões de registros, oferecendo uma interface gráfica abrangente para o usuário, tutoriais de ajuda e visualizações que orientam os usuários em diferentes aspectos durante o processo de anonimização. (SILVA, 2019, p. 35).

Outro *software* importante é o Amnesia *Anonymization*, criado com base nas linguagens de programação Java e JavaScript. No processo de anonimização dos dados o usuário define suas escolhas no anonimato, mas a ferramenta possui uma limitação em relação ao ARX, pois utiliza apenas duas técnicas que podem ser selecionados, o k-anonimato e o km-anonimato. (TERROVITIS, 2017).

De acordo com Terrovitis (2017), o *software* Amnesia é de fácil utilidade pois sua interface gráfica é amigável, funcional e orienta nas várias etapas do processo de anonimato. Além disso, pode ser operada em navegador *web* e aplicativo *desktop* assim os dados importados para anonimização posteriormente são armazenados

localmente e não no *backend* do *software*. Sendo assim o *front-end* do Amnesia oferece:

- A capacidade de carregar os campos de regras do conjunto de dados;
- A capacidade de salvar arquivos anônimos;
- Gerar automaticamente hierarquias de generalização sendo possível a edição;
- Executar um algoritmo de anonimato e explorar a saída;
- Explorar a qualidade dos dados anônimos visualizando os resultados;
- Impor regras de supressão;
- Comparar lado a lado o conjunto de dados original e anônimo;
- Exportar o conjunto de dados anônimo assim como as suas regras de anonimato utilizadas.

O *software* Amnesia é um projeto resultante do financiamento da União Europeia (UE) *OpenAIRE* na qual é disponibilizado gratuitamente para comunidade, podendo ser utilizado com a finalidade de reduzir os riscos em caso de violações de dados além de atender as normas estabelecidas por leis de proteção de dados pessoais. (CRUTZEN; PETERS; MONDSCHEIN, 2019).

2.4 Risco de reidentificação dos dados

De acordo com GPDP (2019), o processo de anonimização de dados é alcançado a partir da aplicação das técnicas de anonimização. No entanto, quanto mais alto for o grau de anonimização o conjunto de dados se torna menos claro, de modo que o risco de reidentificação seja reduzido.

No processo de anonimização é importante atentar-se para a estimativa do risco de reidentificação dos dados de cada indivíduo. Para que a estimativa aconteça é necessário calcular a frequência que os atributos semi-identificadores se manifestam no conjunto de dados anonimizado. (SILVA, 2019).

A reidentificação dos dados ocorre quando as informações de identificação do indivíduo são descobertas em dados anonimizados. Com a reidentificação em um conjunto de dados os identificadores diretos e indiretos tornam-se conhecidos e o titular passa a ser identificado. (LUBARSKY, 2017).

Segundo Khaled el Eman (2013 apud SILVA, 2019, p. 37), os três cenários que conduzem a estimativa de risco são:

- Cenário do promotor: neste cenário supõe que o invasor tem conhecimento que os dados do titular, alvo do ataque, estão contidos no conjunto de dados.
- Cenário do jornalista: o conhecimento sobre o titular dos dados não é assumido.
- Cenário do profissional de *marketing*: supõe que o atacante não está interessado em reidentificar um alvo em específico, mas pretende atacar um número maior de indivíduos.

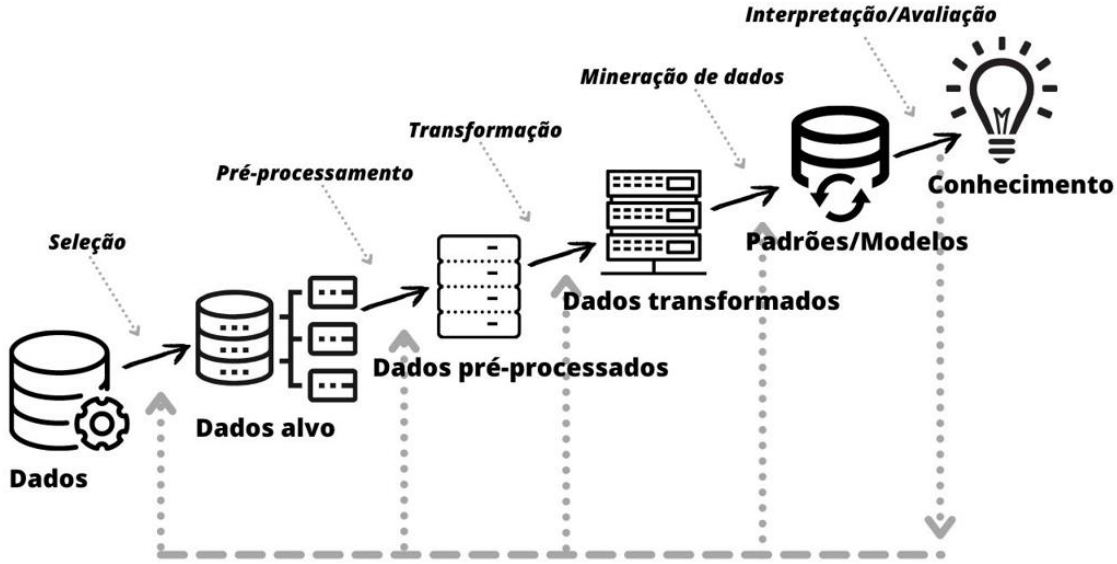
Portanto, com base nos conceitos apresentados, para que um ataque seja efetivado é necessário que uma grande parte dos dados sejam reidentificados. Com um conjunto de dados reidentificados é possível ainda aplicar técnicas de mineração de dados para chegar a outras informações importantes sobre o alvo.

2.5 Mineração de dados

A mineração de dados trata da aplicação de técnicas em um conjunto de dados com a finalidade de extrair ou descobrir informações. O termo *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Banco de Dados é considerado um processo que possui várias etapas, na qual a mineração de dados é uma das etapas essenciais neste processo e consiste na aplicação de técnicas inteligentes com a finalidade de extrair as informações de interesse. (AMO, 2004).

De acordo com Silva (2019) o processo de descobrir as informações por meio dos dados KDD é constituído das seguintes etapas: 1) seleção de dados; 2) pré-processamento; 3) transformação dos dados; 4) mineração de dados; 5) Interpretação e avaliação. As etapas são ilustradas como apresentado na Figura 8.

Figura 8: Processo de descoberta de conhecimento



Fonte: Adaptado de Raval (2012, p. 439).

Conforme Raval (2012, p. 439), o “objetivo final do processo de descoberta de conhecimento e mineração de dados é encontrar os padrões que estão escondidos entre os enormes conjuntos de dados e interpretá-los como conhecimento e informações úteis”. Com base nos conceitos apresentados é importante que seja conceituada algumas técnicas de classificação de dados utilizadas no processo de mineração.

2.5.1 Técnicas de mineração

No processo de mineração dos dados várias técnicas podem ser utilizadas para descobrir informações em bancos de dados e alcançar o resultado pretendido. Nesta seção é apresentada a técnica de classificação dos dados.

2.5.2 Classificação dos dados

Pode-se compreender que “a classificação é o processo de encontrar um conjunto de modelos que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados”. (AMO, 2004, p. 5). Por exemplo, o processo de classificação poderia

ser utilizado por uma empresa para identificar e classificar as pessoas adimplentes e inadimplentes.

Gonçalves (2007) relata que o classificador é utilizado para observar a base de dados através de suas características e assim identificar a classe que os dados pertencem. No processo de classificação duas fases devem ser consideradas:

- Fase de aprendizado: em um conjunto de dados de treinamento aplica-se algum algoritmo classificador.
- Fase de teste: após a aplicação do algoritmo classificador, a etapa de teste tem como objetivo avaliar a acurácia a partir do conjunto de dados teste.

Nas palavras de Silva (2019), a utilidade dos dados anonimizados pode ser melhor entendida a partir da avaliação da acurácia que é a combinação da precisão e exatidão que indicam o número de previsões corretas entre todas as outras previsões realizadas pelo classificador. A acurácia é obtida a partir da matriz de confusão, na qual os erros e exatidão são identificados na predição de classe e o cálculo da acurácia pode ser apurado a partir do número de acertos dividido pelo número de registros testados.

O conjunto de treinamento refere-se a um subconjunto de observações que foram selecionadas aleatoriamente a partir da base de dados a ser analisada. Assim, cada observação no conjunto de treinamento é representada por dois atributos: 1) atributo classe: indica a classe à qual a observação pertence; 2) atributo preditivo: os valores são analisados a fim de descobrir como se relaciona com atributo classe. (GONÇALVES, 2007).

Com base na Tabela 9, o conjunto de dados de treinamento será exemplificado. Neste exemplo tem-se dados de pessoas, observando o atributo classe que é “estabilidade financeira”. Este atributo é utilizado para indicar se uma pessoa possui um conforto financeiro a partir da análise dos seus atributos preditivos que são “escolaridade” e “idade”. Assim a partir de um algoritmo classificador os atributos são classificados e o resultado apresenta a acurácia do conjunto de dados.

Tabela 9: Conjunto de dados de treinamento

NOME	ESCOLARIDADE (atributo preditivo)	IDADE (atributo preditivo)	ESTABILIDADE FINANCEIRA (atributo classe)
Ana	Bacharel	20-60	Não
Silvia	Doutorado	>=60	Sim
João	Mestrado	>=60	Sim
Gustavo	Bacharel	20-60	Não
Leonardo	Bacharel	20-60	Sim
Gabriel	Mestrado	>=60	Sim

Fonte: Elaborado pelo autor, com base em Gonçalves (2007, p.2)

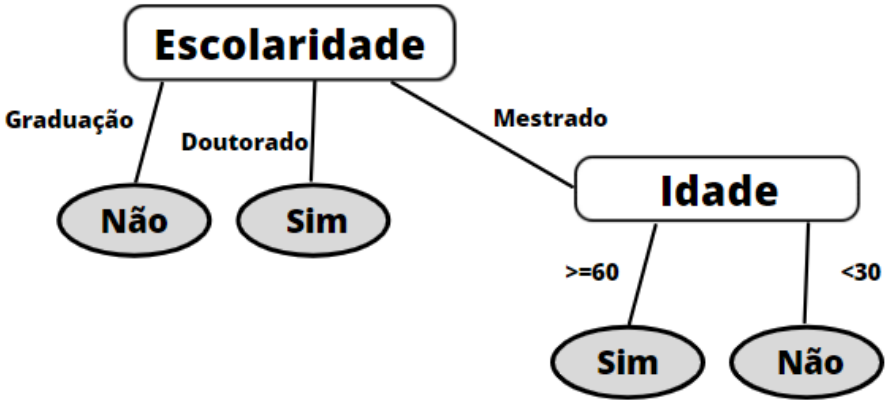
Em vista disso é importante destacar que existem outras técnicas de mineração que podem ser utilizadas com base no tipo de informação que se deseja buscar. No entanto nesta pesquisa o enfoque é no método classificador, o qual possui vários algoritmos que podem ser utilizados tais como as árvores neurais, máquina de vetores de suporte e entre outros.

2.5.3 Árvore de decisão

Como dito existem vários algoritmos de classificação disponíveis para serem utilizados para mineração dos dados, porém o algoritmo da árvore de decisão é utilizado nesta pesquisa. Desse modo, é necessário que conceitos relativos a este algoritmo de classificação sejam evidenciados.

Gonçalves (2007) caracteriza que a árvore de decisão é estruturada por regras de classificação, na qual os caminhos da raiz até a folha são representados por uma das regras. No entanto a árvore de decisão deve ter apenas um caminho da raiz até a folha a partir das observações realizadas no conjunto de dados. A Figura 9 apresenta a estrutura de uma árvore de decisão com base nos atributos da Tabela 9, que indicam se uma pessoa possui estabilidade financeira a partir dos atributos preditivos.

Figura 9: Árvore de decisão com base na tabela 9



Fonte: Adaptado de Gonçalves (2007, p.2)

Um processo importante na construção da árvore é a separação dos dados em conjuntos de treinamento e teste em que a maior parte dos dados é utilizada para treinamento e a menor parte é usada para teste. Sendo assim, o conjunto de treinamento são dados retirados do conjunto de dados original para construção da árvore de decisão em que o conjunto de teste é utilizado para testar o desempenho e precisão da árvore de decisão construída. (MARIN; LOPES, 2013).

De acordo com Castro e Ferrari (2016), o processo de construção da árvore de decisão também conhecido como indução de árvores de decisão é utilizado para determinar a classe de um objeto a partir dos valores dos atributos. Marin e Lopes (2013) afirmam que a seleção dos atributos é realizada de acordo com critérios estatísticos que buscam os atributos mais relevantes, ou seja, os algoritmos de indução buscam dividir os dados de um nó pai com o objetivo de minimizar o grau de impureza dos nós filhos. A entropia é a medida de seleção de atributos que define a pureza na qual os valores baixos indicam mais pureza.

O cálculo da entropia é definido conforme apresentado na equação (2.1), em que Cj é a quantidade de amostras da classe e S a quantidade total das amostras.

$$\text{Info(S)} = \text{entropia(S)} = -\sum_{j=1}^{k} \left(\frac{Cj}{S}\right) * \log 2 \left(\frac{Cj}{S}\right) \qquad (2.1)$$

Quanto maior for os resultados da entropia, maior será a desordem dos dados necessitando mais esforço do algoritmo para organizar os dados a suas classes pertencentes. Sendo assim a entropia pode ajudar a decidir qual atributo do conjunto de dados é o melhor a ser utilizado para criação de partições. Quanto mais variabilidade de classes presentes em um conjunto de dados mais impura será as classes. Assim uma partição pura é aquela que contém exemplares de uma única classe. (SILVA; PERES; BOSCARIOLI, 2016).

> Se todas as medidas de entropia resultarem em 0, todas as partições serão consideradas puras, e saberemos que o atributo A é bom para ser usado como critério para particionar os dados, já que as partições que ele gera possuem, cada uma, exemplares de uma única classe. Na indução de uma árvore de decisão, o uso desse atributo levaria à finalização da construção do modelo classificador. Se algumas das medidas de entropia resultarem em valores maiores que 0, é sinal de que as partições contêm exemplares de classes diferentes. (SILVA; PERES; BOSCARIOLI, 2016, p. 101).

No entanto é importante que a análise de informação necessária seja aplicada de forma a obter a quantidade de esforço necessário para chegar as partições puras. A equação de informação necessária é dada pela fórmula (2.2) em que: S_i: é a quantidade de amostras para a partição; S: quantidade total das amostras; m: quantidade de partições e info(S_i): entropia total para a partição.

$$\text{info(S,A)} = \sum_{i=1}^{m} \left(\frac{Si}{S}\right) * info\ (S_i) \qquad (2.2)$$

A partir do conceito de informação necessária o conceito ganho de informação é estabelecido. E assim o atributo que possui o maior ganho de informação esperado é selecionado pelo ganho máximo, ou seja, seleciona o atributo que possui menor tamanho esperado pelas subárvores, admitindo que a raiz é o principal nó. (MARIN; LOPES, 2013).

A equação do ganho de informação é dada pela fórmula (2.3). Esse cálculo consiste na subtração da entropia de todo o conjunto pela entropia de cada atributo.

$$ganho(S,A) = info(S) – info(S,A) \quad (2.3)$$

A árvore completa é o modelo de classificação construído para predizer resultados e contribuir nas tomadas de decisões. As medidas normalmente utilizadas para a avaliação de classificadores é a acurácia ou taxa de classificações corretas. A partir da acurácia é possível analisar a classificação dos registros classificados corretamente e incorretamente. (SILVA; PERES; BOSCARIOLI, 2016).

De acordo com Silva, Peres e Boscarioli outra forma de realizar a análise da árvore de decisão é por meio da matriz de confusão que possui dimensões n x n, em que n é o número de classes presentes e as repostas corretas do classificador geram os valores na diagonal principal da matriz.

A matriz de confusão possui significados em cada uma de suas células na qual pode indicar problemas maiores ou menores nos resultados do classificador. (SILVA; PERES; BOSCARIOLI, 2016). O Quadro 2 apresenta os significados das células da matriz de confusão para o problema binário.

Quadro 2: Matriz de confusão de classificação binária

	Positivo	Negativo
Positivo	Verdadeiros positivos	Falsos negativos
Negativo	Falsos positivos	Verdadeiros negativos

Fonte: Elaborado pelo autor com base em Silva, Peres e Boscarioli (2016, p. 137).

Nas palavras de Silva, Peres e Boscarioli (2016) os significados são caracterizados como:

- Verdadeiros positivos indicam que o exemplar pertence à classe positiva e que o classificador realizou a classificação como pertencente a classe positiva.
- Falso positivos indicam que o exemplar pertence à classe negativa e o classificador o classificou como pertencente a classe positiva.
- Verdadeiros negativos indicam que o exemplar pertence à classe negativa, porém o classificador realizou a classificação pertencente a classe negativa.
- Falsos negativos indicam que o exemplar pertence à classe positiva, porém o classificador realizou a classificação pertencente a classe negativa.

A classificação baseada em árvore de decisão com o algoritmo C4.5 possui opções padrão para sua construção. Entre as mais úteis e comuns, estão a opção -C 0.25 que é um valor que afeta a poda da árvore de decisão e valores baixos causam uma poda mais brusca em relação aos valores altos. Essa opção padrão funciona bem para a maioria dos experimentos. Já a opção mínimo padrão -M 2 é utilizada para evitar árvores estranhas com pouco poder preditivo. O algoritmo C4.5 exige que qualquer teste utilizado na árvore tenha pelo menos dois resultados com um número mínimo de casos desse tipo. O valor padrão é 2 e pode ser modificado para valores mais elevados em casos com muitos dados ruidosos presentes. (QUINLAN, 1993).

Conforme descrito por Saravanan e Gayathri (2018, p. 188), a “árvore de decisão pode ser construída de forma moderadamente rápida em comparação com outros métodos de classificação”. Dessa maneira, diante dos conceitos evidenciados sobre o algoritmo árvore de decisão é importante apresentar o *software* utilizado para realizar a mineração de dados a partir de algoritmos e técnicas.

2.6 A ferramenta WEKA

O WEKA é um *software open source* utilizado para tarefas relacionadas a mineração de dados. Este *software* possui várias ferramentas para preparar os dados e aplicar técnicas de mineração, como por exemplo a técnica de classificação dos dados, entre outras.

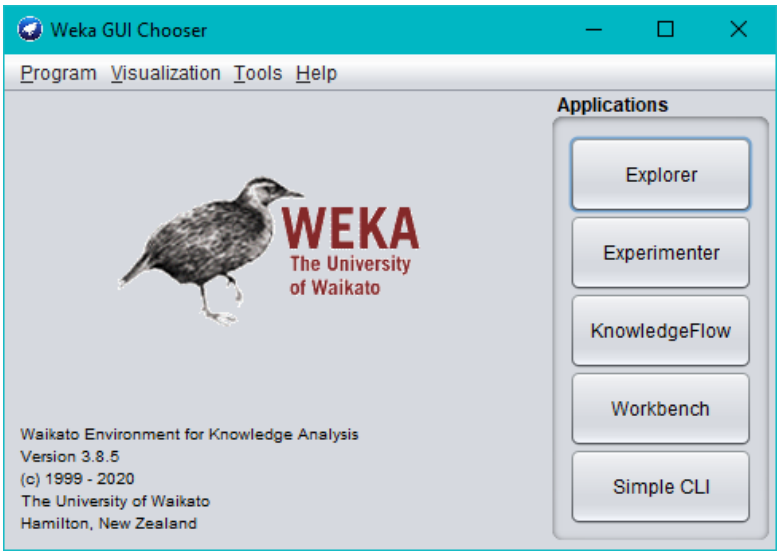
Segundo Witten et al. (2016), a ferramenta WEKA pode ser utilizada de três maneiras, sendo elas:

- Aplicando o método de aprendizagem em um conjunto de dados para analisar as saídas com objetivo de saber mais sobre os dados;
- Gerando previsões sobre os modelos aprendidos;
- Aplicando vários modelos de aprendizagem de modo a comparar os desempenhos para escolher qual previsão usar.

Witten et al. (2016) caracteriza que a maneira mais fácil de utilizar o WEKA é por sua interface gráfica que possui cinco menus para acesso a distintas interfaces que são: a) *Explorer*: nessa interface se pode ler um arquivo de conjunto de dados e a partir dele construir uma árvore de decisão e visualizar os resultados; b)

Experimenter: o usuário pode utilizar vários algoritmos e comparar os resultados de modo a escolher o melhor algoritmo para seu conjunto de dados; c) KnowledgeFlow: possui as mesmas funções do *explorer* porém, essa interface permite que as configurações sejam ajustadas para processamento de dados em fluxo; d) *Workbench*: essa interface é a combinação com todas as interfaces anteriormente citadas, formando uma única interface; e) *Simple CLI:* essa interface funciona por códigos de comandos suportados pelo WEKA. Tais menus são apresentados na Figura 10.

Figura 10: Tela inicial do *software* WEKA



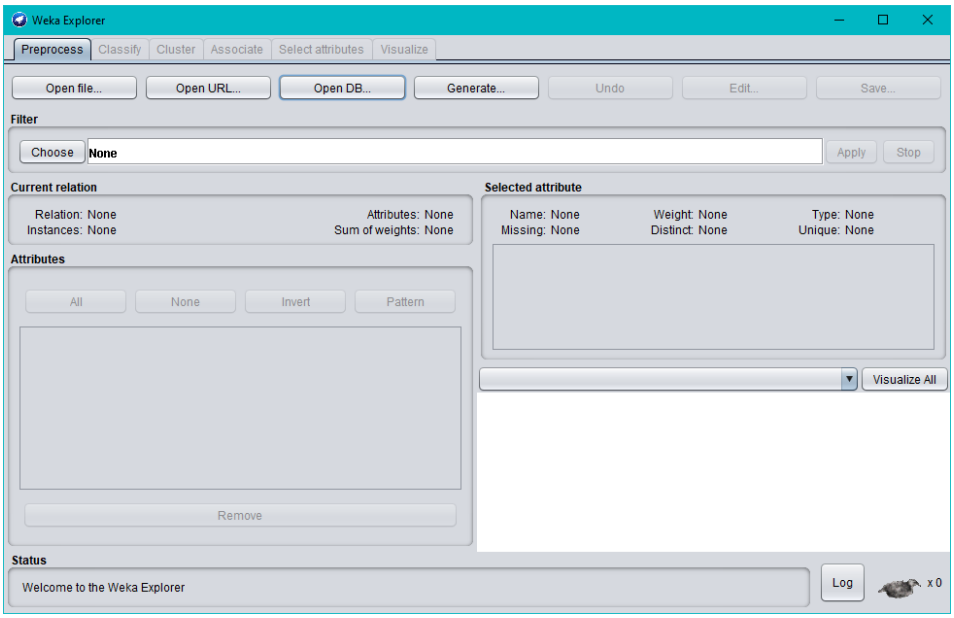
Fonte: Elaborado pelo autor.

Os dados carregados são apresentados para o usuário em uma planilha. Estes dados podem ser importados de arquivos que estão no formato *Atribute Relation File Format* (ARFF) que é o padrão do WEKA e nos formatos *Comma Separated Values* (CSV), *JavaScript Object Notation* (JSON), C4.5 e binário. É possível ainda conectar com banco de dados por conexão *Java Database Connectivity* (JDBC) apenas informando o *Uniform Resource Locator* (URL), usuário e senha, para que assim seja realizado o acesso para experimentos em banco de dados. É relevante destacar a interface *explorer*, pois a partir desta é possível aplicar algoritmos de classificação e regressão, como árvore de decisão (J48), baseada no algoritmo C4.5, dentre outros no conjunto de dados importado. (WITTEN et al., 2016).

O algoritmo J48 pode ser utilizado para executar resultados precisos a partir da classificação de vários dados. Assim o algoritmo cria uma árvore de decisão que analisa as informações e cria subconjuntos menores para se basear em uma decisão. (SARAVANAN; GAVATHRI, 2018).

A Figura 11 apresenta a interface *explorer* utilizada para minerar os dados a partir de suas técnicas.

Figura 11: Interface explorer



Fonte: Elaborado pelo autor.

A interface do *explorer* permite criar experimentos de grande escala, executá-los e quando finalizados possibilita que as estatísticas geradas sejam analisadas, o que é observado ser bastante utilizado em algoritmos de mineração de dados e trabalhos sobre mineração de dados.

2.7 Trabalho correlato

Há vários trabalhos importantes disponíveis relacionados a anonimização, mineração e privacidade dos dados. O trabalho de Silva (2019), tem como objetivo propor uma abordagem para anonimização dos dados em plataformas de análise de dados de forma que o processo de anonimização e mineração dos dados permita a

extração de conhecimento dos dados anonimizados para tomadas de decisão de modo que a privacidade das pessoas seja protegida.

Um conjunto de dados com 480 mil transações de cartões com dados de usuários de ônibus foi utilizado em seus experimentos. Na anonimização dos dados as técnicas de supressão e generalização foram aplicadas. A mineração de dados aconteceu com a aplicação dos seguintes algoritmos classificadores:

- Zero R;
- LWL+Zero R;
- SDG e *Navies Bayes*.

Os algoritmos classificadores foram aplicados sobre o atributo Tipo de Cartão de Crédito (Visa ou Mastercard) em que o classificador avalia o tipo de cartão de crédito utilizado pelos usuários dos ônibus. A mineração de dados ocorreu com os dados originais e dados anonimizados com análises sobre os impactos que anonimização dos dados provocaria nos algoritmos de classificação.

Os resultados das acurácias obtidas não podem ser considerados bons para realizar tomadas de decisões, visto que a acurácia do conjunto de dados sem anonimização com a aplicação de todos os algoritmos classificadores obteve como resultado uma acurácia de 51% para os classificadores Zero R, SDG e *Naives Bayes* e 69% com os classificadores LWL+Zero R. Estes resultados com o conjunto de dados original se apresentaram ruins, dificultando na tomada de decisões. Com a aplicação das técnicas de anonimização as acurácias se apresentaram piores.

3 MATERIAIS E MÉTODOS

Esse capítulo tem por objetivo especificar os materiais e métodos utilizados no desenvolvimento deste trabalho, além de suas etapas em relação as atividades realizadas, de modo que seja possível compreender os processos que auxiliaram na obtenção dos resultados.

3.1 Métodos

De acordo com Gil (2018) é possível definir uma pesquisa como um procedimento que tem o objetivo de proporcionar respostas a problemas propostos, ou seja, a pesquisa é requerida quando não se tem informações suficientes para responder um problema. Assim a pesquisa pode ser realizada com desejo de conhecer algo de maneira mais eficiente ou eficaz.

A pesquisa em seu contexto científico pode ser classificada de acordo com a sua natureza, objetivos ou procedimentos técnicos. Este trabalho, quanto à natureza é classificado como trabalho original no qual busca conhecimento novo com base em observações e teorias construídas para compreender e explicá-las. (WAZLAWICK, 2014).

Quanto aos seus objetivos, este trabalho pode ser classificado como uma pesquisa exploratória e descritiva, pois inicialmente o propósito foi de criar uma maior familiaridade com a LGPD, anonimização e mineração dos dados a partir do levantamento bibliográfico. Em um segundo momento, são aplicadas técnicas de anonimização e mineração de dados em um conjunto de dados de crédito disponibilizado pela ferramenta WEKA com o propósito de comparar as acurácias obtidas nestes processos, e então descrever os resultados obtidos. Essa pesquisa é quantitativa pois os resultados são apresentados em termos numéricos e qualitativa em que os resultados são comparados de forma descrita. (GIL, 2018).

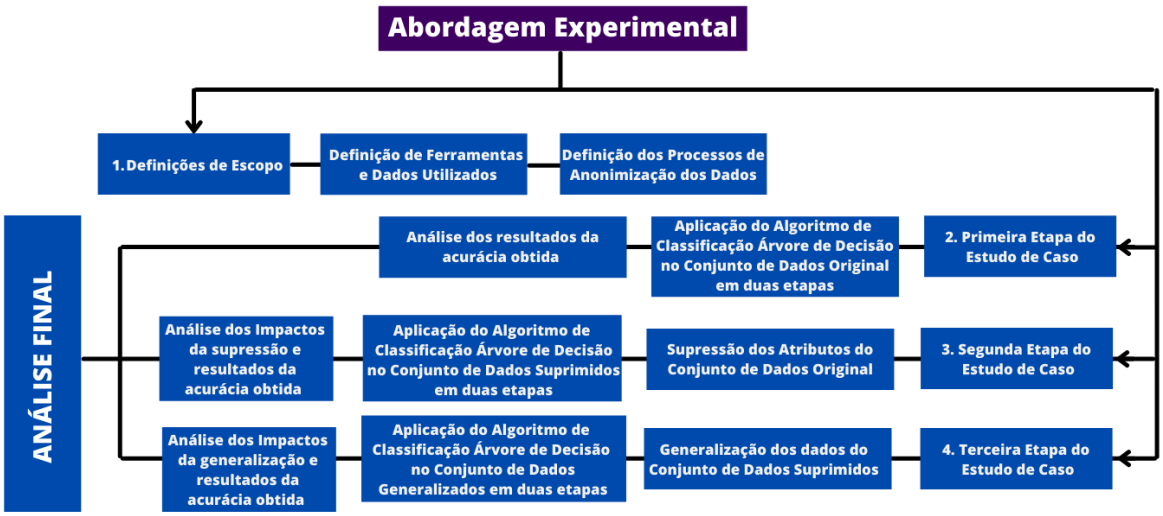
Em relação aos procedimentos técnicos, a pesquisa pode ser classificada como pesquisa bibliográfica e experimental, sendo que a pesquisa bibliográfica foi elaborada a partir de estudo de artigos, livros, teses e monografias, com o objetivo de obter conhecimento já escrito sobre LGPD, anonimização e mineração de dados. Além disto, a manipulação de variáveis, as quais podem provocar alterações nas pesquisas,

de tal modo que é possível observar se as intervenções realizadas produzem os resultados esperados, essa é uma caracterização de uma pesquisa experimental. Sendo assim, neste trabalho são manipuladas variáveis do conjunto de dados de crédito com o objetivo de analisar os resultados obtidos da mineração de dados aplicada em cada um dos conjuntos de dados até se obter um conjunto de dados anonimizados. (WAZLAWICK, 2014).

3.1.1 Abordagem experimental

Com o objetivo de fornecer uma visão geral sobre as etapas de implementação e dos experimentos realizados, cada uma das etapas desta abordagem é apresentada e descrita conforme mostra Figura 12.

Figura 12: Abordagem experimental



Fonte: Elaborado pelo autor com base em Silva (2019, p.63).

Definições de Escopo: Nesta etapa definiu-se a abordagem de anonimização técnicas, ferramentas e o conjunto de dados utilizado nos experimentos.

Primeira Etapa do Estudo de Caso – Aplicação do Algoritmo de Classificação Árvore de Decisão no Conjunto de Dados Original em duas etapas: o propósito dessa etapa é avaliar a acurácia do conjunto de dados original em que o algoritmo de classificação árvore de decisão utilizado na mineração de dados é aplicado em duas

etapas. Na primeira etapa o algoritmo classificador árvore de decisão é executado de modo que todos os dados do conjunto de dados original são utilizados como treinamento e teste. Na segunda etapa utiliza-se do mesmo algoritmo classificador e o conjunto de dados original é dividido em 66% dos dados para treinamento e o restante para teste.

Segunda Etapa do Estudo de Caso – Supressão dos Atributos do Conjunto de Dados Original com Mineração dos Dados em duas etapas: o objetivo desta etapa é suprimir os atributos do conjunto de dados original, a remoção dos atributos é realizada visto que não são necessários. A supressão dos atributos é executada pelo *software* Amnesia que tem por objetivo transformar dados pessoais em dados anônimos. A mineração de dados em duas etapas realizada na primeira etapa do estudo de caso também é realizada com o conjunto de dados suprimidos. Ao final da primeira e segunda etapa do estudo de caso, devem ser analisados os impactos que a supressão dos atributos provoca no algoritmo de classificação árvore de decisão, ou seja, as acurácias do conjunto de dados original e suprimidos precisam ser observadas.

Terceira Etapa do Estudo de Caso – Generalização dos Dados do Conjunto de Dados Original com Mineração dos Dados em duas etapas: esta etapa tem como objetivo generalizar os dados dos atributos semi-identificadores do conjunto de dados suprimidos com a finalidade de chegar à abordagem proposta que é obter um conjunto de dados anonimizados (suprimido e generalizado). O experimento com a mineração de dados em duas etapas deve ser executado no conjunto de dados anonimizados assim como acontece na primeira e segunda etapa do estudo de caso. Com as três etapas finalizadas as acurácias devem ser comparadas. As comparações dos resultados das acurácias se iniciam pelo conjunto de dados original até o conjunto de dados anonimizados com objetivo de obter respostas para as questões de pesquisa.

3.2 Materiais

Neste trabalho são utilizados o conjunto de dados de crédito disponibilizado pela ferramenta WEKA, que foi utilizada para realizar a mineração dos dados, o *software* Amnesia *Anonymization tool* e um notebook da marca Samsung com as seguintes especificações:

- Modelo Samsung *Expert* NP300E5M-XD1BR;
- Windows 10 *Home*;
- Intel Core i5-7200U de 7ª Geração (2.5 GHz até 3.1 GHz) 3 MB L3 Cache;
- HD 1 TB;
- 8 GB RAM.

3.2.1 Conjunto de dados utilizados

Para realizar este experimento é utilizado o conjunto de dados de crédito, provenientes do repositório *data* da ferramenta WEKA. As características destes dados são detalhadas a seguir.

3.2.2 Dados de crédito

Para realizar o experimento foi utilizado o conjunto de dados de crédito alemão que possui 1000 registros contendo informações de pessoas devedoras que são classificadas como adimplentes ou inadimplentes. O conjunto de dados possui 20 atributos com as respectivas informações de cada um dos indivíduos, desconsiderando o atributo classificação. Este conjunto de dados de crédito apresenta atributos semi-identificadores que podem ser utilizados para reidentificação do titular dos dados.

As informações disponibilizadas possuem dados pessoais dos clientes que solicitaram crédito. No entanto, as técnicas de supressão e generalização dos dados são utilizadas com o objetivo de anonimizar o conjunto de dados. Este conjunto de dados foi escolhido por possuir dados pessoais, assim os atributos poderiam ser anonimizados para iniciar os experimentos. Outro fator que foi levado em consideração para a utilização deste conjunto de dados é que a maioria dos conjuntos de dados pesquisados não possuíam dados pessoais dos usuários. Porém, vale ressaltar que para realizar o experimento, poderia ter sido adicionado registros de usuários, ou seja, dados pessoais fictícios de usuários no conjunto de dados original.

O Quadro 3 apresenta as técnicas de anonimização aplicadas aos principais atributos do conjunto de dados de crédito.

Quadro 3: Técnicas de anonimização aplicadas aos atributos

Atributos da tabela	Tipo de dados	Técnica utilizada
Status da conta corrente	*	Supressão
Duração no mês	*	Supressão
Histórico de crédito	*	Supressão
Propósito do crédito	Semi-identificador	Supressão
Quantidade de crédito	Semi-identificador	Generalização
Estado da poupança	*	Supressão
Tempo de emprego	*	Manter dados
Compromisso de parcelamento	*	Manter dados
Sexo e status pessoal	*	Manter dados
Outros fiadores e devedores	*	Manter dados
Residência desde	*	Manter dados
Propriedade	*	Manter dados
Idade	Semi-identificador	Generalização
Outros planos de parcelamento	*	Supressão
Casa	*	Manter dados
Possui telefone	*	Supressão
Créditos existentes neste banco	*	Manter dados
Trabalho	*	Manter dados
Número de dependentes	*	Manter dados
Possui telefone	*	Manter dados
Trabalhador estrangeiro	Semi-identificador	Supressão

Fonte: Elaborado pelo autor com base em Silva (2019, p. 70).

A anonimização dos dados foi aplicada nos atributos semi-identificadores a partir das técnicas de supressão e generalização. A supressão ocorreu em atributos semi-identificadores e atributos considerados não úteis para tomada de decisão. Nos

atributos semi-identificadores necessários, foram aplicadas as técnicas de generalização dos dados.

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentadas as discussões e resultados que foram alcançados no decorrer deste trabalho. Em todos os testes realizados nos conjuntos de dados de crédito, o modelo de privacidade k-anonimato é aplicado. Nos atributos semi-identificadores utilizou-se as técnicas de supressão e generalização a fim de alcançar o anonimato dos dados, segurança da informação, privacidade dos dados e a conformidade com a LGPD.

4.1 Experimento com dados originais

Os experimentos foram realizados a partir de um conjunto de dados no qual o arquivo está nomeado como credit-g no formato ARFF, que se encontra no diretório de instalação do *software* WEKA na pasta ...\Weka-3-8-5\data. Neste arquivo os atributos e dados foram traduzidos da língua inglês para o português do Brasil.

4.1.1 Experimento 1 - *Use training set*

No primeiro experimento com a mineração de dados no conjunto de dados original foi utilizado a técnica de classificação dos dados e o algoritmo árvore de decisão J48 com os seguintes parâmetros: -C 0.25 -M 2 com o propósito de classificar se as pessoas estão adimplentes ou inadimplentes. No teste de árvore de decisão foi utilizado a opção *Use training set*. Essa opção de teste utiliza 100% dos dados de treinamento como teste.

A partir da mineração dos dados realizada no conjunto de dados original, obteve-se um resultado com uma acurácia de 85,5%, em que 855 registros foram classificados corretamente e 145 registros incorretamente, conforme apresentado na Figura 13.

Figura 13: Resultado da árvore de decisão – *use training set*

```
=== Summary ===

Correctly Classified Instances         855               85.5    %
Incorrectly Classified Instances       145               14.5    %
Kappa statistic                          0.6251
Mean absolute error                      0.2312
Root mean squared error                  0.34
Relative absolute error                 55.0377 %
Root relative squared error             74.2015 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,956    0,380    0,854      0,956    0,902      0,640    0,857     0,905     bom
               0,620    0,044    0,857      0,620    0,720      0,640    0,857     0,783     mau
Weighted Avg.  0,855    0,279    0,855      0,855    0,847      0,640    0,857     0,869

=== Confusion Matrix ===

   a   b   <-- classified as
 669  31 |   a = bom
 114 186 |   b = mau
```

Fonte: Elaborado pelo autor.

Esse experimento gerou de uma árvore com 140 nós, em que 103 são nós folhas. Os resultados desse experimento mostraram uma acurácia muito boa com a utilização da árvore de decisão.

4.1.2 Experimento 2 - *Percentage split*

No segundo experimento foi utilizado a árvore de decisão e a opção que divide o conjunto de dados original em 66% para treinamento e 34% para testes que é a opção de teste *Percentage split*. Neste experimento foi gerado uma árvore com 140 nós na qual 103 são nós folhas. A Figura 14 apresenta os resultados do experimento em que é possível observar que a metodologia adotada obteve uma acurácia de 72,65%, na qual 247 registros foram classificados corretamente e 93 incorretamente.

Figura 14: Resultados árvore de decisão - *percentage split*

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         247               72.6471 %
Incorrectly Classified Instances        93               27.3529 %
Kappa statistic                          0.2687
Mean absolute error                      0.3351
Root mean squared error                  0.4836
Relative absolute error                 80.8066 %
Root relative squared error            108.7935 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,836    0,578    0,801      0,836    0,818      0,270    0,605     0,782     bom
                 0,422    0,164    0,481      0,422    0,450      0,270    0,605     0,370     mau
Weighted Avg.    0,726    0,468    0,716      0,726    0,721      0,270    0,605     0,673

=== Confusion Matrix ===

   a   b   <-- classified as
 209  41 |   a = bom
  52  38 |   b = mau
```

Fonte: Elaborado pelo autor.

Analisando o primeiro experimento é possível observar que este apresentou uma melhor acurácia com a utilização de todo o conjunto de dados como treinamento e teste. Porém, essa metodologia que utiliza 100% dos dados de treinamento como teste não é a mais adequada pois o modelo de treinamento conhece os dados de teste. Já o segundo experimento se mostrou com um acurácia inferior, porém pode ser considerado um modelo com resultados mais aceitáveis para a realidade. Como o objetivo dos experimentos é analisar a acurácia do conjunto de dados original em relação a acurácia do conjunto de dados anonimizados, para realizar esse procedimento a supressão e a generalização dos dados é executada no conjunto de dados original.

Com base na LGPD o tratamento deste conjunto de dados original relacionados ao crédito pode ser realizado utilizando a base legal de proteção de crédito. Na lei não é especificado como os dados devem ser tratados por quem utiliza dessa base legal. No caso, o compartilhamento de dados é uma das operações relacionadas ao tratamento de dados. No entanto, ao utilizar a base legal de proteção de crédito e compartilhar todas as informações é como se estivesse indo na contramão da LGPD.

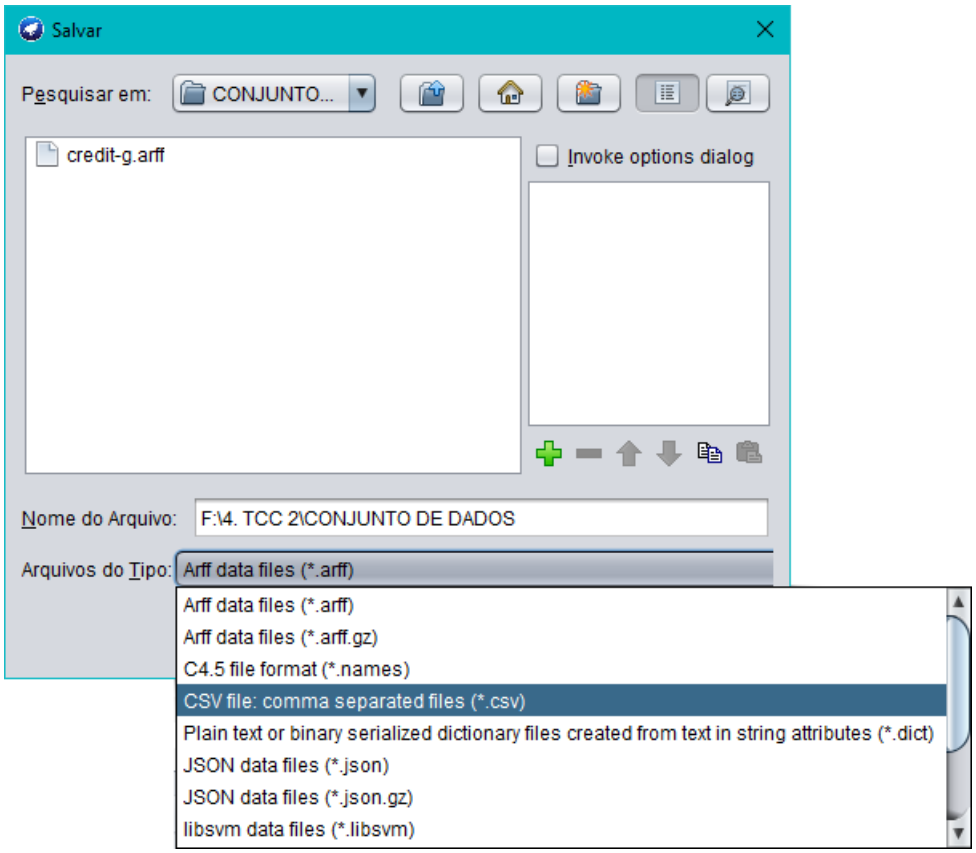
Com base na análise da mineração de dados realizada no conjunto de dados original, o resultado mostrou-se bom nos dois experimentos realizados, pois obteve uma acurácia acima dos 70% para ambos. Já a operação de compartilhamento dos dados utilizando a base legal de proteção ao crédito gera alguns contratempos em relação a lei, mas é a base legal que pode ser utilizada por empresas de crédito para realizar o tratamento dos dados.

4.2 Experimento com supressão dos atributos

Para este experimento é realizada a supressão de alguns atributos do conjunto de dados original utilizado no experimento anterior. A supressão dos atributos foi realizada utilizando o *software* Amnesia. Essa remoção dos atributos acontece uma vez que estes atributos não são necessários no conjunto de dados anonimizados.

Para realizar a supressão dos atributos no *software* Amnesia, o conjunto de dados original no formato de arquivo ARFF foi convertido para o formato CSV. Antes de iniciar o experimento com a supressão dos atributos no Amnesia é necessário que na ferramenta WEKA na janela do *Preprocess* o conjunto de dados original seja salvo no formato de arquivo CSV. Para isso ao clicar no botão *save* em Arquivos do Tipo selecione o formato do arquivo como “CSV *file: comma separated files (*.csv)*” define-se o nome do arquivo e o diretório para salvar conforme apresentado na Figura 15.

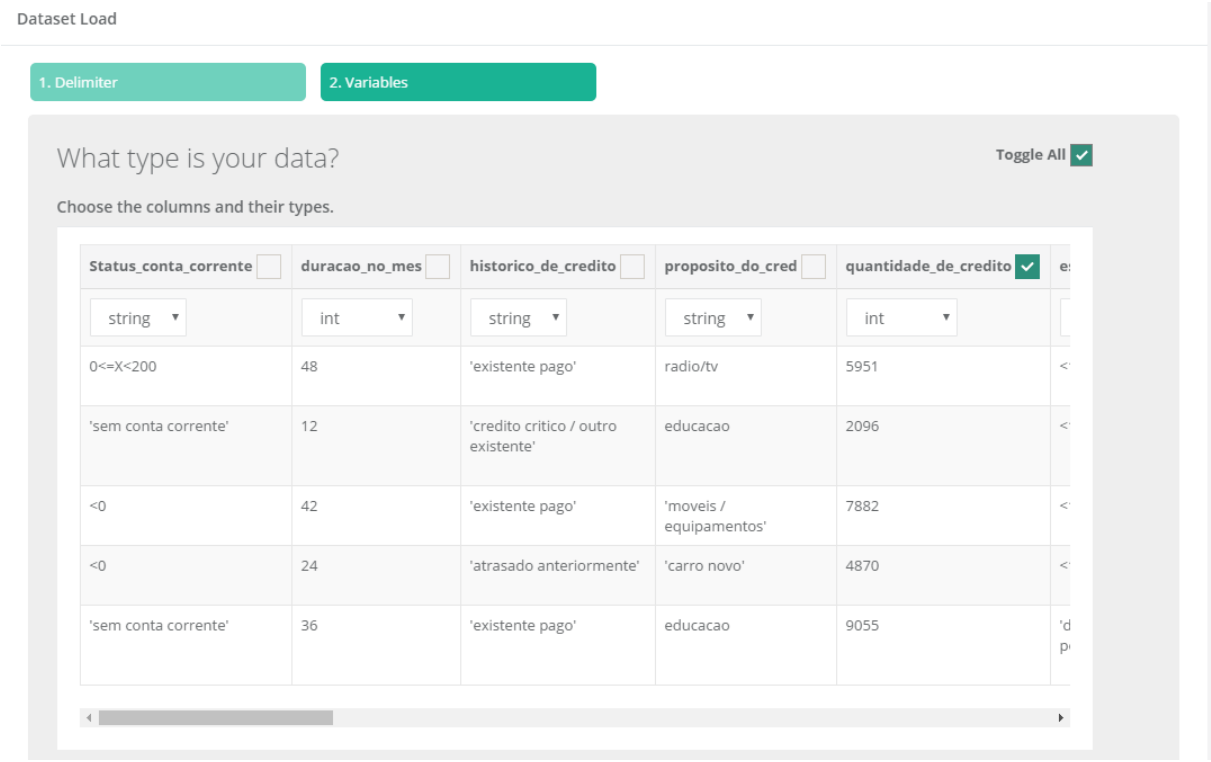
Figura 15: Convertendo arquivo ARFF para CSV no WEKA



Fonte: Elaborado pelo autor.

Ao importar o arquivo com os dados para o Amnesia o conjunto de dados original é carregado e os atributos e dados são apresentados em uma tabela de dados com carregamento automático na qual é possível definir o tipo dos atributos. Caso o carregamento automático esteja incoerente quanto ao tipo dos atributos é possível definir o tipo do atributo. Na mesma janela para realizar a supressão dos atributos listados anteriormente é preciso desmarcar as *checkbox* que se encontram ao lado dos atributos, conforme mostra a Figura 16.

Figura 16: Supressão de atributos no Amnesia



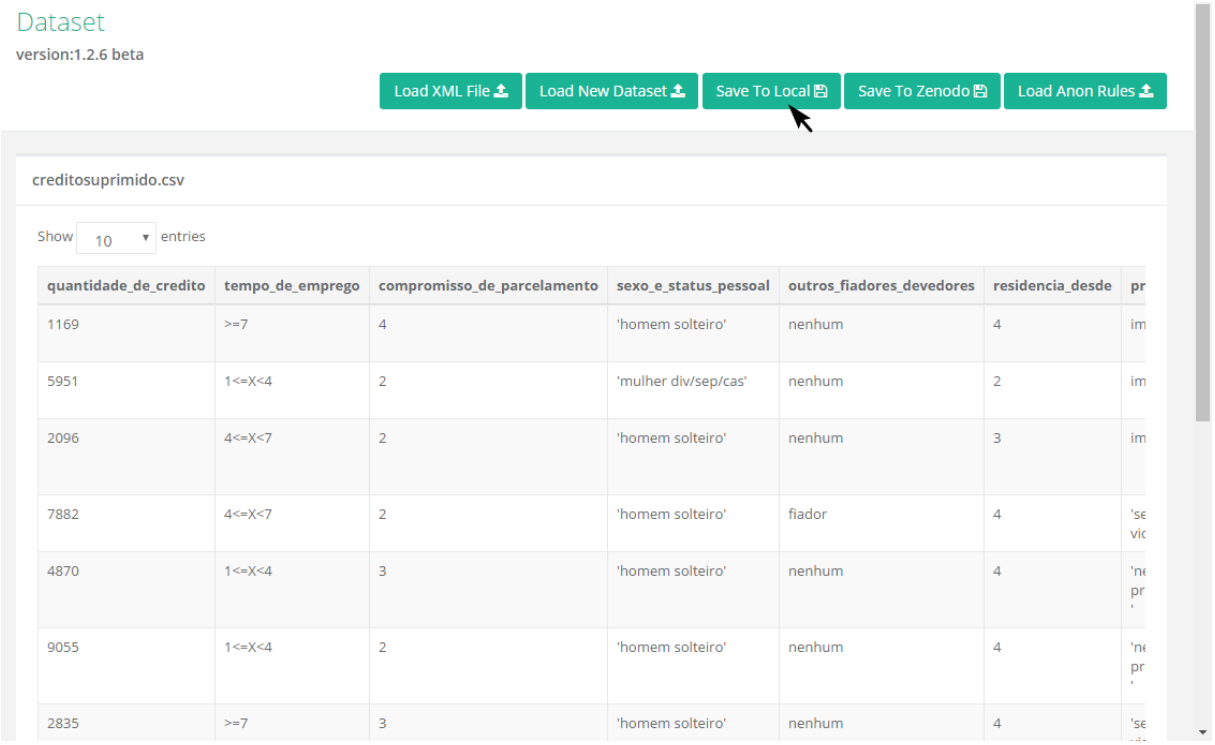
Fonte: Elaborado pelo autor.

Os atributos suprimidos do conjunto de dados original foram:

- *Status* conta corrente;
- Duração no mês;
- Histórico de crédito;
- Proposito do crédito;
- Estado de poupança;
- Outros planos de parcelamento;
- Possui telefone;
- Trabalhador estrangeiro.

Ao desmarcar as checkbox e finalizar a supressão, uma nova tabela de dados com os atributos removidos é apresentada conforme a Figura 17. No botão *Save To Local* o arquivo com conjunto de dados suprimido é salvo no formato CSV para que o terceiro e quarto experimento relacionado a mineração de dados com conjunto de dados suprimidos sejam realizados na ferramenta WEKA.

Figura 17: Conjunto de dados original suprimido



Fonte: Elaborado pelo autor.

4.2.1 Experimento 3 - *Use training set*

Para os experimentos com conjunto de dados suprimidos, o arquivo no formato CSV é importado no WEKA. Para executar os testes também foi utilizado o algoritmo classificador J48 com as opções -C 0.25 -M 2. Assim como anteriormente nos dois experimentos relacionados a mineração de dados também foi utilizado as opções de testes *Use training set* e *Percentage split*. Com a supressão de 8 dos 21 atributos do conjunto de dados original apenas 13 atributos e seus respectivos dados são carregados no WEKA. A Figura 18 ilustra os resultados da árvore de decisão com a opção de teste *Use training set*.

Figura 18: Resultado árvore de decisão – *use training set*

```
Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         784               78.4    %
Incorrectly Classified Instances       216               21.6    %
Kappa statistic                          0.3946
Mean absolute error                      0.3253
Root mean squared error                  0.4033
Relative absolute error                 77.4342 %
Root relative squared error             88.0134 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,959     0,623     0,782      0,959     0,861      0,440    0,716     0,821     bom
               0,377     0,041     0,796      0,377     0,511      0,440    0,716     0,584     mau
Weighted Avg.  0,784     0,449     0,786      0,784     0,756      0,440    0,716     0,750

=== Confusion Matrix ===

   a   b   <-- classified as
 671  29 |   a = bom
 187 113 |   b = mau
```

Fonte: Elaborado pelo autor.

Analisando a Figura 18, observa-se que o experimento com conjunto de dados original utilizando a mesma opção de teste apresentou melhores resultados que nos experimentos com conjunto de dados suprimidos. O experimento com atributos removidos mostrou-se ter 78,40% de acurácia em que 784 registros foram classificados corretamente e 216 incorretamente. Uma árvore foi gerada com 102 nós em que 65 são nós folhas.

Com a supressão dos atributos e dados em função da necessidade da anonimização dos dados é possível observar que a acurácia do conjunto dados suprimidos obteve acurácia inferior ao conjunto de dados original. Embora a acurácia tenha sido inferior devido a supressão, este resultado pode ser considerado bom.

4.2.2 Experimento 4 - *Percentage split*

No quarto experimento utilizando a árvore de decisão os registros do conjunto de dados suprimido são divididos em 66% para treinamento e o restante para testes. Analisando a Figura 19 pode-se observar que este experimento mostrou-se ter

68,82% de acurácia pois conseguiu classificar corretamente 234 registros e 106 registros incorretamente. O experimento gerou uma árvore com 102 nós em que 65 são nós folhas. No primeiro experimento no qual foi utilizado 100% do conjunto de dados suprimidos para treinamento e testes os resultados se apresentaram melhores.

Figura 19: Resultado árvore de decisão - *percentage split*

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         234               68.8235 %
Incorrectly Classified Instances       106               31.1765 %
Kappa statistic                          0.1035
Mean absolute error                      0.3745
Root mean squared error                  0.4918
Relative absolute error                 90.3125 %
Root relative squared error            110.646  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,848    0,756    0,757      0,848    0,800      0,107    0,569     0,756     bom
                 0,244    0,152    0,367      0,244    0,293      0,107    0,569     0,327     mau
Weighted Avg.    0,688    0,596    0,654      0,688    0,666      0,107    0,569     0,642

=== Confusion Matrix ===

   a   b   <-- classified as
 212  38 |   a = bom
  68  22 |   b = mau
```

Fonte: Elaborado pelo autor.

Com base na LGPD o compartilhamento dos dados deste conjunto de dados suprimidos pode ser realizado entre empresas de crédito utilizando a base legal de proteção de crédito. Porém, os dados não podem ser compartilhados publicamente na internet ou com terceiros apenas com a supressão da forma que se encontra, pois os atributos semi-identificadores se apresentam em evidência, gerando assim riscos de reidentificação dos dados. Neste sentido é necessário aplicar a técnica de generalização dos dados para tornar mais ainda os dados anônimos.

4.3 Experimento com generalização dos dados

A partir do conjunto de dados suprimido os experimentos com a generalização dos dados são iniciados com intuito de obter uma maior anonimização dos dados. Para realizar a generalização dos dados é utilizado o modelo de privacidade k-anonimato do *software* Amnesia.

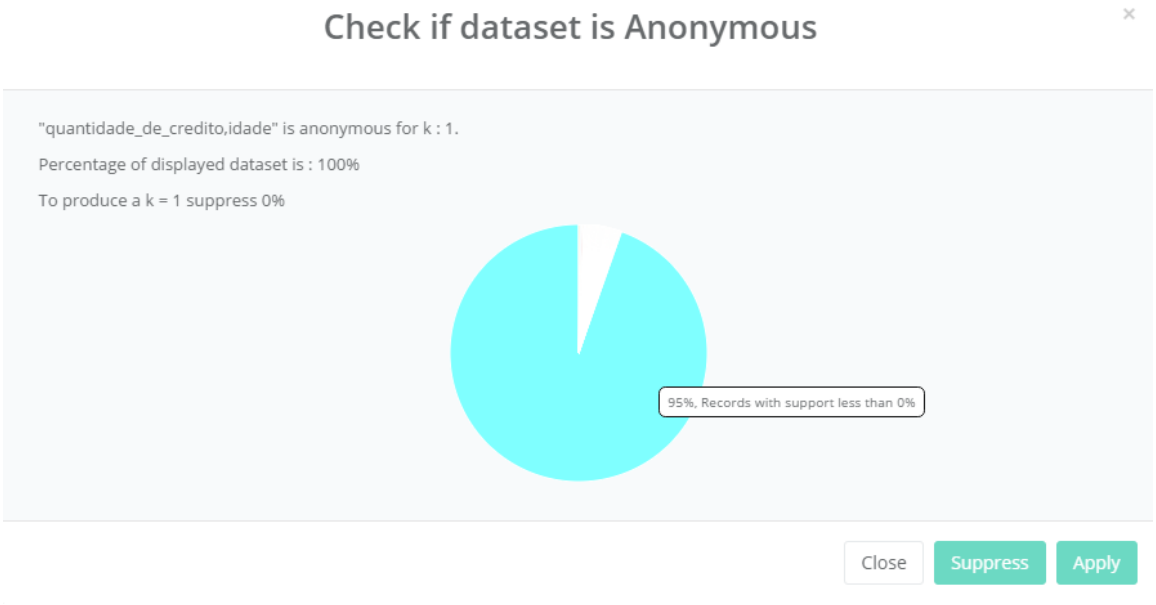
Neste experimento com conjunto de dados suprimidos, o arquivo com os dados importados no Amnesia são utilizados para realizar a generalização dos dados nos atributos semi-identificadores. Os atributos semi-identificadores são:

- Quantidade de crédito;
- Idade.

No conjunto de dados estes atributos possuem informações sobre a quantidade de crédito fornecida ao titular dos dados, além de apresentar a idade do indivíduo. De forma a evitar a reidentificação dos dados por meio destes atributos os dados são generalizados.

Com conjunto de dados suprimido carregado em uma tabela que apresenta os atributos e seus respectivos dados, ao clicar no botão *check anonymization* é possível verificar se o conjunto de dados está anonimizado de acordo com o k-anonimato. Assim, o Amnesia inicia a verificação que se deve informar os atributos e o valor de k-anonimato. Com base no conjunto de dados suprimido, somente os atributos quantidade de crédito e idade são selecionados e o valor de k foi igual a 1. Com resultado da verificação do anonimato para k igual a 1 é possível observar que os dados dos dois atributos se mostraram anônimos, não sendo necessário suprimir nenhum dado do conjunto de dados, conforme apresenta a Figura 20.

Figura 20: Resultado da verificação do anonimato dos dados dos atributos quantidade de crédito e idade do conjunto de dados para k=1.



Fonte: Elaborado pelo autor.

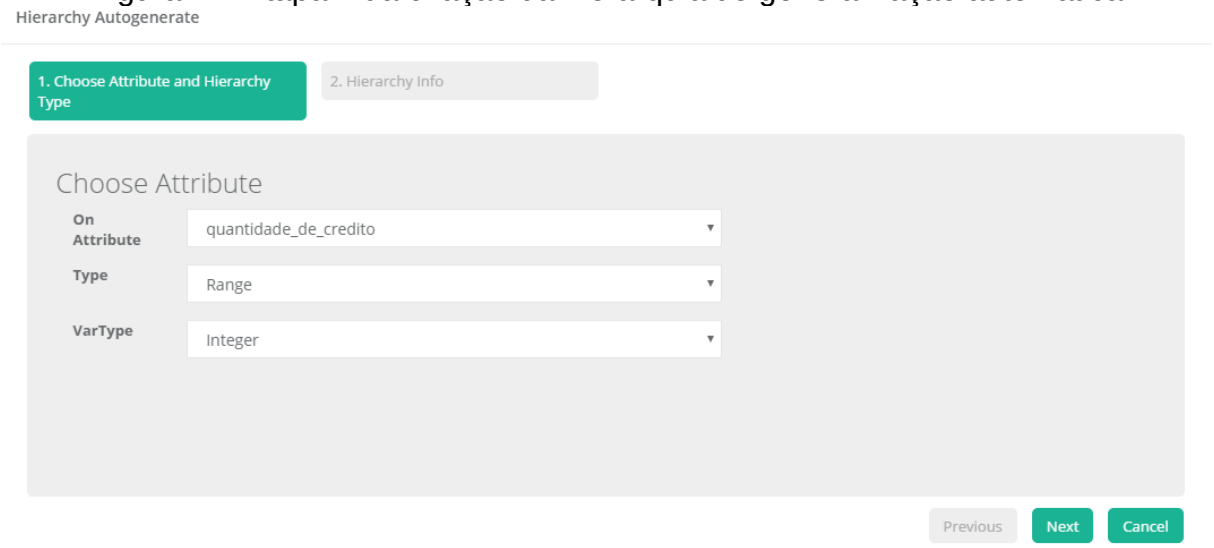
É importante destacar que uma solução insegura que viole o k-anonimato pode ser apresentada nos dados e ao invés de generalizar ainda mais os dados, o encarregado de realizar o tratamento dos dados pode optar por suprimir estes dados. Observe que na Figura 20 a opção “Supress”, poderia ser utilizada no atributo idade caso possuísse dados que violassem o k-anonimato, assim o gráfico com as estatísticas apresentaria a porcentagem de dados que violaram a garantia de privacidade dos dados, sendo possível escolher suprimir os dados a fim de transformar uma solução insegura em segura.

As próximas etapas do experimento é anonimizar os dados com a generalização dos atributos na qual o k utilizado será igual a 1. Após a verificação do anonimato dos dados segue-se para a etapa de aplicação das hierarquias de generalização. Na janela de hierarquias de generalização no botão *Autogenerate Hierarchy* uma hierarquia de geração automática deve ser criada para generalizar os dados dos atributos definidos.

Os atributos quantidade de crédito e idade possuem dados do tipo inteiro, ou seja, valores numéricos inteiros. Na etapa 1 da hierarquia de generalização automática, ao selecionar o atributo quantidade de crédito, na caixa de seleção *type* a opção *range* é escolhida com a finalidade de generalizar os valores. A Figura 21

mostra a etapa 1 da hierarquia de generalização e o mesmo procedimento acontece para o atributo idade.

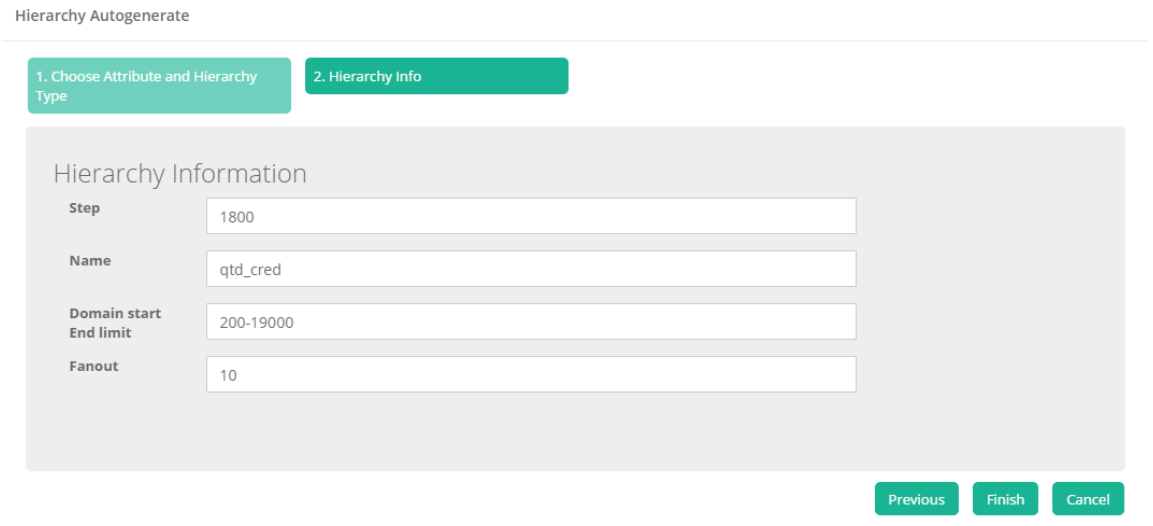
Figura 21: Etapa 1 da criação da hierarquia de generalização automática



Fonte: Elaborado pelo autor.

Na etapa 2 da hierarquia de generalização os valores são definidos automaticamente em conformidade com os dados. Porém, é possível definir os intervalos para que o algoritmo de anonimização atenda a privacidade a ser atingida. Com base no atributo quantidade de crédito os valores do domínio, nome da hierarquia e *fanout* são definidos conforme os valores apresentados na Figura 22.

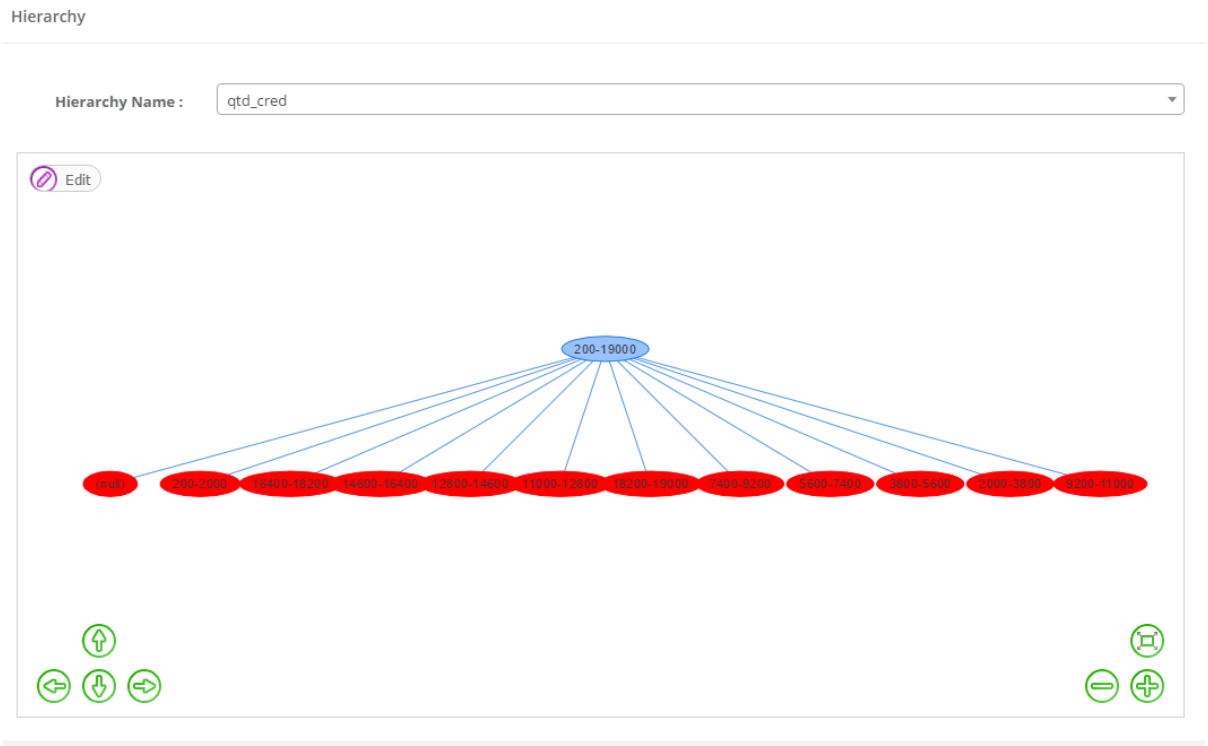
Figura 22: Definindo os valores da hierarquia de generalização



Fonte: Elaborado pelo autor.

Os valores da hierarquia de generalização realizados na etapa 2 foram personalizados. Ao finalizar as etapas da hierarquia de generalização uma árvore é gerada com vários nós no qual todos são conduzidos a um único nó raiz. A quantidade de nós folhas é definido conforme o valor de *fanout* inserido na etapa 2. A Figura 23 apresenta a árvore de hierarquia gerada conforme os valores definidos na Figura 22.

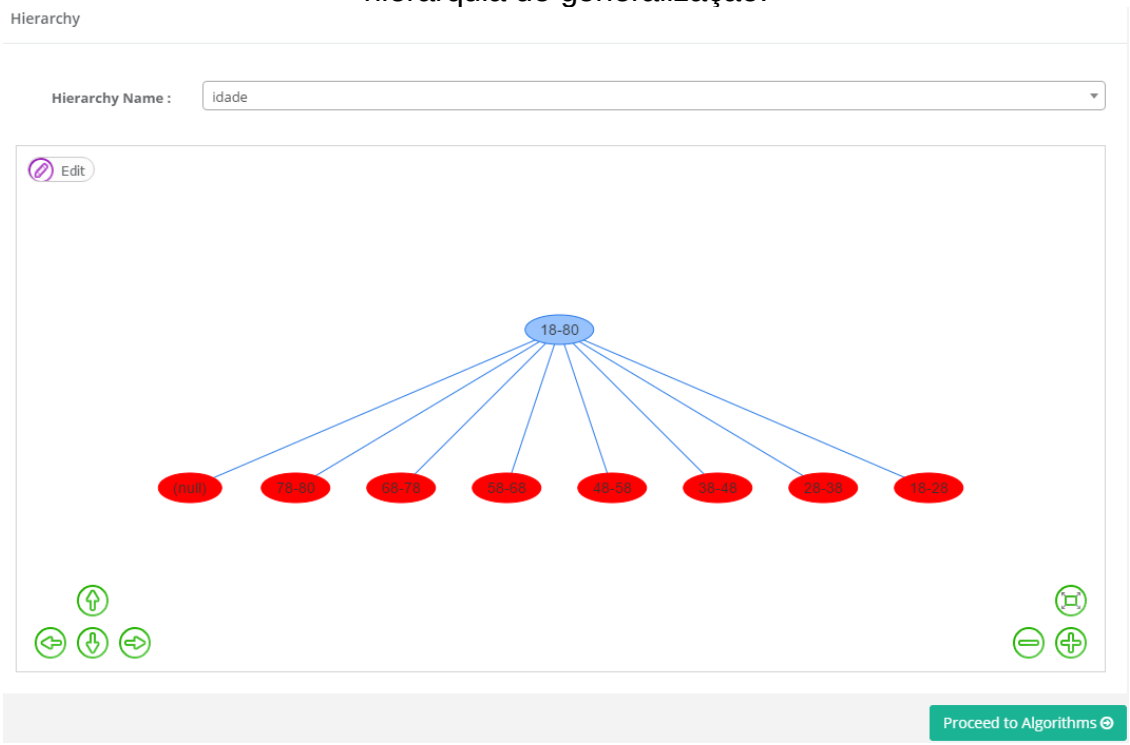
Figura 23: Árvore do atributo quantidade de crédito gerada com base nos parâmetros definidos na hierarquia de generalização.



Fonte: Elaborado pelo autor.

Em relação ao atributo idade na etapa 2 da hierarquia de generalização os valores inseridos nos seguintes campos foram: Intervalo do domínio de 18-80; Faixa entre os intervalos do domínio de 10; *Fanout* com 10 que define a quantidade de nós filhos de cada nó. A Figura 24 mostra o resultado da árvore gerada a partir da hierarquia de generalização.

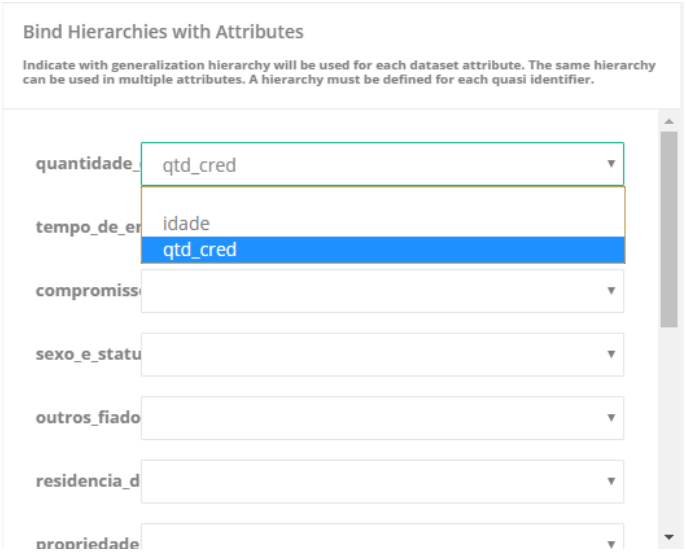
Figura 24: Árvore do atributo idade gerada com base nos parâmetros definidos na hierarquia de generalização.



Fonte: Elaborado pelo autor.

Na etapa 3 a execução do algoritmo é realizada por meio da tela *Algorithms*, na qual as hierarquias são vinculadas aos seus respectivos atributos. Observe na Figura 25 a aplicação das hierarquias nos atributos.

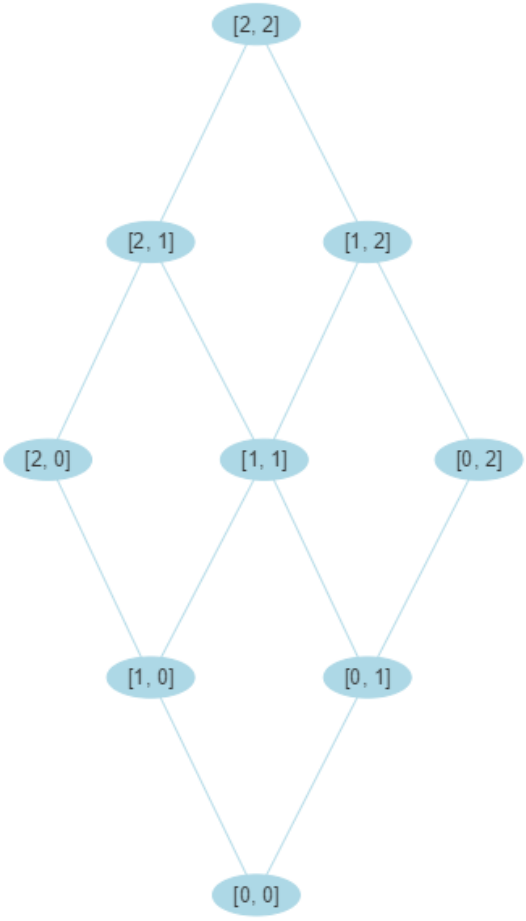
Figura 25: Vinculando hierarquias de generalização aos seus respectivos atributos



Fonte: Elaborado pelo autor.

Ao finalizar a vinculação das hierarquias com os atributos e executar o algoritmo, um gráfico de solução é apresentado. Assim, a partir do gráfico de solução gerado pode-se analisar que os nós azuis apresentam soluções de k-anonimato para cada um dos atributos. Todas as soluções geradas indicam uma solução segura, em que é possível selecionar a solução de forma que o conjunto de dados anônimo seja gerado. É importante destacar que na verificação do anonimato realizada nos atributos e dados no início deste experimento se mostraram anônimos para k igual a 1. A Figura 26 apresenta as soluções de k-anonimato na qual a solução do gráfico aplicada é a (1,1), ou seja, todos os dois atributos com k igual a 1.

Figura 26: Gráfico de solução de k-anonimato



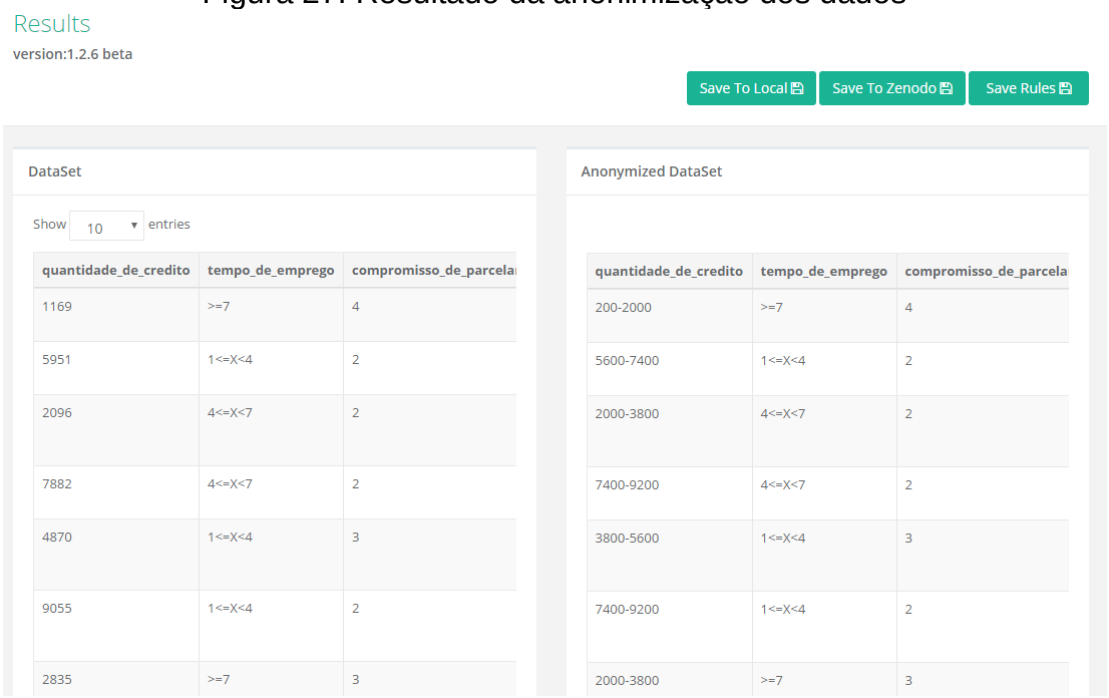
Fonte: Elaborado pelo autor.

É possível observar na Figura 26, que a árvore de soluções de k-anonimato apresenta soluções seguras (nós azuis). Para todas as soluções nenhum nó vermelho foi exibido, os casos que apresentam soluções não seguras (nós vermelhos) é aquele

em que os dados descumprem com a privacidade dos dados, sendo assim se por acaso o encarregado de realizar a anonimização escolher uma solução não segura, este deve utilizar a supressão para remover os dados que colaboram com a reidentificação dos dados, caso contrário os dados não removidos podem contribuir com a reidentificação dos dados.

Com a aplicação do nó (1,1) uma tela com os resultados é apresentada, na qual é possível comparar o conjunto de dados original suprimido com o conjunto de dados anonimizados, conforme mostra a Figura 27. Para que os experimentos relacionados a mineração dos dados sejam executados o arquivo com conjunto de dados anonimizados deve ser salvo no diretório local no formato CSV.

Figura 27: Resultado da anonimização dos dados



Fonte: Elaborado pelo autor.

4.3.1 Experimento 5 - *Use training set*

Para realizar o quinto experimento de mineração de dados o conjunto de dados anonimizados foi gerado no formato de arquivo CSV. O arquivo com os dados anonimizados foram realizados como nos experimentos anteriores e como resultados 748 pessoas estão classificadas como adimplentes e 252 como inadimplentes utilizando a opção *Use training set*. Para este experimento também foi utilizado a

técnica de classificação composta pelo algoritmo árvore de decisão com as mesmas opções. Os testes foram iniciados utilizando o conjunto de dados anonimizados na qual 100% dos dados foram utilizados para treinamento e teste, conforme apresenta a Figura 28.

Figura 28: Resultados árvore de decisão – *Use training set*

```
=== Summary ===

Correctly Classified Instances         748               74.8    %
Incorrectly Classified Instances       252               25.2    %
Kappa statistic                          0.25
Mean absolute error                      0.3692
Root mean squared error                  0.4297
Relative absolute error                 87.883  %
Root relative squared error             93.7637 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,970    0,770    0,746      0,970    0,843      0,320    0,642     0,777     bom
                 0,230    0,030    0,767      0,230    0,354      0,320    0,642     0,483     mau
Weighted Avg.    0,748    0,548    0,752      0,748    0,697      0,320    0,642     0,689

=== Confusion Matrix ===

   a   b   <-- classified as
 679  21 |   a = bom
 231  69 |   b = mau
```

Fonte: Elaborado pelo autor.

O conjunto de dados original que foi anonimizado a partir da supressão e generalização dos dados obteve um resultado que se mostrou adequado, pois obteve 74,8% de acurácia o que pode ser considerado bom em relação aos experimentos anteriores. A partir do conjunto de dados original é possível observar que até se chegar no conjunto de dados anonimizados a acurácia foi reduzida em 10,7%. Essa redução ocorre devido a supressão dos atributos e generalização dos dados do conjunto de dados original. Porém, mesmo com essa redução é possível utilizar os resultados para tomar decisões sem comprometer a identificação do titular dos dados.

4.3.2 Experimento 6 - *Percentage split*

No sexto experimento foi utilizado a árvore de decisão e a opção *Percentage split* que divide o conjunto de dados anonimizados em 66% para treinamento e o restante para teste. Esse experimento gerou uma árvore com 140 nós na qual 103 são nós folhas. Na Figura 29 é apresentado os resultados obtidos na qual o modelo obteve 72,6% de acurácia em que 247 registros foram classificados corretamente e 93 registros incorretamente.

Figura 29: Resultado árvore de decisão - *percentage split*

```
=== Summary ===

Correctly Classified Instances         247               72.6471 %
Incorrectly Classified Instances        93               27.3529 %
Kappa statistic                          0.2687
Mean absolute error                      0.3351
Root mean squared error                  0.4836
Relative absolute error                 80.8066 %
Root relative squared error            108.7935 %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,836    0,578    0,801      0,836    0,818      0,270    0,605     0,782     bom
                 0,422    0,164    0,481      0,422    0,450      0,270    0,605     0,370     mau
Weighted Avg.    0,726    0,468    0,716      0,726    0,721      0,270    0,605     0,673

=== Confusion Matrix ===

   a   b   <-- classified as
 209  41 |   a = bom
  52  38 |   b = mau
```

Fonte: Elaborado pelo autor.

O quinto experimento com conjunto de dados anonimizados se apresentou com uma melhor acurácia. Porém, o sexto experimento apresentou resultados que pode ser considerado mais aplicáveis em experimentos reais, pois este não utiliza os mesmos dados para treinamento e testes, logo os dados se ajustam o melhor possível aos dados de testes.

Com base na LGPD, como os dados foram anonimizados e perderam sua vinculação direta com o titular dos dados a empresa que trata os dados se protege das sanções e normas aplicadas pela lei. Com o conjunto de dados anonimizados é possível compartilhar os dados com terceiros desde que não seja possível reidentificar os dados.

É importante destacar que a anonimização dos dados é apenas uma das estratégias para a gestão de riscos. Essa estratégia aperfeiçoa a segurança da informação dentro da empresa, além de gerar confiança nos serviços prestados. No entanto não é importante somente anonimizar os dados, é preciso que os ativos da organização estejam protegidos no intuito de evitar os incidentes de segurança.

A situação em que o conjunto de dados original se encontre armazenado em discos rígidos de servidores internos ou serviços de armazenamento em nuvem é comum. Sendo assim é importante que a segurança da informação esteja presente para garantir que os ativos utilizados para armazenar o conjunto de dados original se encontrem protegidos contra ameaças e vulnerabilidades. O conjunto de dados anonimizados garante a privacidade das pessoas de tal forma que o compartilhamento dos dados aconteça sem causar danos para os titulares e controladores dos dados.

5 CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um enfoque no processo de anonimização de dados em que se utilizou as técnicas e modelos de anonimização de materiais já publicados, que combinadas podem equilibrar o compromisso entre a privacidade e utilidade dos dados. Para cada uma das etapas da aplicação das técnicas de anonimização dos dados realizadas com o conjunto de dados de crédito a mineração de dados obteve os resultados das acurácias. Foram realizadas análises para se obter respostas do impacto que a anonimização dos dados provocou nos resultados das acurácias obtidas na mineração de dados. Este enfoque na anonimização dos dados visa contribuir com a segurança da informação, gestão de riscos e cumprimento da LGPD. Em todas as etapas referentes ao tratamento dos dados realizadas no experimento, sugestões com base na LGPD foram apresentadas.

Com base nos resultados apresentados no estudo de caso pode-se concluir que as técnicas de anonimização (supressão e generalização) combinadas com a mineração podem contribuir na utilidade dos dados e como consequência aumentar a proteção dos dados pessoais. Assim a perda de informação causada pela anonimização dos dados pode proporcionar resultados sobre a acurácia que podem ser consideráveis de tal forma que a privacidade dos dados não seja comprometida e seja possível tomar decisões, desde que a anonimização e mineração seja realizada com prudência pelo encarregado de realizar o tratamento dos dados. Além disto, a anonimização dos dados evita a reidentificação dos titulares dos dados desde que aplicada corretamente.

Este trabalho também se submeteu a responder as seguintes questões de pesquisa que justificam o tema e as contribuições do trabalho.

Q1. A anonimização dos dados contribui ou agrava a qualidade dos resultados dos algoritmos de classificação utilizados no processo de mineração dos dados?

A anonimização e mineração dos dados foram executadas nas duas últimas etapas (supressão e generalização) do experimento. Com a anonimização realizada em etapas foi possível avaliar os impactos que a anonimização proporcionou nos resultados do algoritmo de classificação (árvore de decisão J48) utilizado na mineração. Em análise aos experimentos realizados pode-se afirmar que a aplicação das técnicas de anonimização (supressão e generalização) não causaram

consequências relevantes na acurácia e na qualidade do algoritmo classificador. Comparando a acurácia da segunda e terceira etapa do experimento em que foi utilizado a opção *percentage split* é possível observar que a acurácia da terceira etapa em que foi aplicado as técnicas de generalização de dados no conjunto de dados suprimidos se apresentou melhor que a acurácia da segunda etapa na qual foi utilizado somente a técnica de supressão. A avaliação destas etapas, comprovou que a anonimização dos dados não causa grandes consequências na acurácia e no desempenho do algoritmo classificador.

Com o aumento da acurácia na terceira etapa do experimento, foi possível observar que este aumento ocorreu devido a aplicação da técnica de generalização dos dados nos atributos. Sendo assim quando os atributos com os dados generalizados passam pelo algoritmo classificador na mineração de dados a acurácia tende a crescer e a perda de informações é resultante desse processo. Esse aumento no resultado da acurácia acontece visto que a generalização diminui a frequência dos dados no conjunto de dados.

A técnica de supressão dos dados foi aplicada em atributos insignificativos e em alguns atributos semi-identificadores. No entanto é importante que a supressão seja aplicada cuidadosamente em observância aos resultados da acurácia, de tal modo que os resultados sejam úteis ao propósito.

Em relação ao k-anonimato modelo de privacidade utilizado pelo *software* Amnesia, os dados que não atendem aos critérios do modelo devem ser removidos do conjunto de dados. Assim deve haver vários indivíduos no conjunto de dados para que essa remoção não aconteça. No entanto, ao utilizar a técnica de generalização do k-anonimato em que o k seja alto, a probabilidade de reidentificação dos dados se torna menor e como os valores da classe atributo diminuem por conta da generalização a acurácia sofre uma falsa melhoria como visto na terceira etapa do experimento.

Portanto, percebe-se que a anonimização dos dados se executada de forma correta pode contribuir com o processo de mineração de dados, sendo assim possível tomar decisões sem comprometer a privacidade das pessoas.

Q2. A anonimização dos dados contribui com segurança da informação, gestão de riscos e cumprimento da LGPD?

A anonimização dos dados realizada no estudo de caso com conjunto de dados de crédito permitiu compreender que a anonimização garante a privacidade dos dados. Nos experimentos realizados com o conjunto de dados original as técnicas de anonimização dos dados foram aplicadas, gerando um conjunto de dados anonimizado com intuito de compartilhar os dados com terceiros sem comprometer a privacidade das pessoas e cumprir com a LGPD.

De acordo com experimentos executados é possível afirmar que a anonimização dos dados tende a contribuir com a segurança da informação em relação ao ativo (conjunto de dados original). Sendo assim ao realizar o tratamento dos dados é importante que o responsável adote medidas de segurança para proteger os dados pessoais de acessos não autorizados. É relevante destacar que a segurança da informação tem por objetivo cumprir com a integridade, confidencialidade, disponibilidade, legalidade, autenticidade e a auditabilidade. Entretanto é considerável rememorar que um conjunto de dados anonimizados é o resultado da aplicação de técnicas de anonimização em um conjunto de dados original, que se encontra localizado em algum ativo dentro da organização. A fim de cumprir com a segurança da informação é necessário que todos os ativos dentro de uma organização sejam considerados.

Portanto ao adotar a anonimização dos dados em um conjunto de dados, o ativo que armazena o conjunto de dados original deve ser protegido contra vazamentos. Para isso uma gestão e governança da segurança da informação deve ser muito bem avaliada e estruturada para proteção dos ativos. Assim a anonimização dos dados pode ser utilizada em um conjunto de dados como uma forte estratégia na gestão de riscos, contribuindo na proteção do ativo relacionado ao conjunto de dados. Porém é importante frisar que a segurança da informação deve ser também estabelecida nos ativos em que se encontram as informações originais. Não é efetivo anonimizar o conjunto de dados se o ativo em que se encontra os dados originais apresenta-se vulnerável a vazamentos e ataques.

Q3. Considerando os dados utilizados nos experimentos, é possível determinar uma simetria entre a privacidade e a utilidade dos dados?

Ao utilizar as técnicas de supressão e generalização dos dados é possível definir um balanceamento entre utilidade e privacidade dos dados. Por exemplo, se optar por mais privacidade as técnicas de anonimização podem utilizadas com maior

aplicação sobre os dados com valor de k-anonimato elevado, no entanto ao aumentar a anonimização dos dados a utilidade dos dados diminui. O *software* Amnesia auxilia o encarregado dos dados no processo de anonimização. Porém quem decide entre utilidade e privacidade dos dados é o encarregado de realizar o tratamento dos dados, ou seja, aquele que aplica as técnicas de anonimização.

Os resultados das acurácias obtidas nos experimentos com supressão e generalização dos dados mostrou que a aplicação destas técnicas de anonimização não provocou grandes impactos de tal forma que tomada de decisão seja possível sem comprometer a privacidade das pessoas.

Sendo assim com base na execução dos experimentos relacionados a anonimização dos dados é possível afirmar que o equilíbrio entre privacidade e utilidade dos dados pode ser determinado pelo encarregado de realizar o tratamento dos dados, pois este poderá ajustar e definir o valor do k-anonimato desejado ou exigido.

Q4. Com o conjunto de dados anonimizados é possível compartilhar com terceiros sem que haja a reidentificação dos dados?

A anonimização dos dados tem por objetivo a não reidentificação dos dados por isso toda atenção e cuidado ao realizar a anonimização dos dados é essencial. Com base na LGPD um conjunto de dados anonimizado é considerado anonimizado desde que não permita a reidentificação do titular dos dados, além disso o conjunto de dados anonimizado se encontra resguardado da aplicação de multas e sanções. No entanto em caso de reidentificação dos dados as punições da LGPD são aplicáveis.

Os resultados obtidos neste trabalho indicam que a anonimização dos dados é efetiva desde que não proporcione a reidentificação dos dados. Sendo assim, a anonimização dos dados deve ser realizada em observância a perda de informação (avaliação quantitativa). Para compartilhar os dados com terceiros no intuito de que a tomada de decisões seja possível é preciso que na mineração dos dados (avaliação qualitativa) se tenha pouca variação ou melhora da acurácia em relação ao conjunto de dados original. Posto isso um conjunto de dados anonimizados pode ser compartilhado com terceiros sem comprometer a privacidade das pessoas além de ser uma estratégia a ser utilizada na gestão de riscos contribuindo com a segurança da informação e cumprimento da LGPD.

5.1 Trabalhos futuros

Para trabalhos futuros pode-se utilizar diferentes conjuntos de dados reais com intuito de que os experimentos possam ser ampliados empregando de novas técnicas de anonimização dos dados como encobrimento de caracteres, agregação e entre outras. Além disso outros modelos de anonimização podem ser aplicados tais como l-diversidade e t-proximidade. Entretanto para realizar a aplicação destes e outros modelos de anonimização a ferramenta ARX (*open source*) pode ser utilizada, pois está é composta de vários outros modelos de anonimização disponíveis.

Um trabalho futuro de maior proporção pode ser desenvolvido utilizando de ataques de reidetificação no conjunto de dados anonimizados com a tentativa de reidentificar o titular dos dados, para isso os ataques com os cenários promotor, jornalista e profissional de *markenting* podem ser efetuados.

.

REFERÊNCIAS BIBLIOGRÁFICAS

AMO, Sandra. Técnicas de mineração de dados. **ResearchGate**, Uberlândia, MG, UFMG Faculdade de computação, 2004. Disponível: https://www.researchgate.net/profile/Sandra-Amo/publication/260300816_Tecnicas_de_Mineracao_de_Dados/links/54230bd80cf290c9e3ae25e3/Tecnicas-de-Mineracao-de-Dados.pdf. Acesso em: 23 agos. 2021.

BIONI, Bruno et al. **Tratado de proteção de dados pessoais**. 1. ed. Rio de Janeiro: Forense Ltda, 2020.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Redação dada pela Lei nº 13.853, de 2019. Brasília, DF: Senado Federal, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 15 fev. 2021.

BRASIL. **Guia de boas práticas lei geral de proteção de dados (LGPD)**. agos. 2020. Disponível em: https://www.gov.br/governodigital/pt-br/seguranca-e-protecao-de-dados/guias/guia_lgpd.pdf. Acesso em: 10 mar. 2021.

BRITO, Felipe Timbó; MACHADO, Javam. Preservação de privacidade de dados: Fundamentos, Técnicas e Aplicações. **ReserchGate**, Ceara, 1. ed, 2017. E-book. Disponível em: https://www.researchgate.net/profile/Felipe-Brito-4/publication/318726149_Preservacao_de_Privacidade_de_Dados_Fundamentos_Tecnicas_e_Aplicacoes/links/597a3540a6fdcc61bb05b98a/Preservacao-de-Privacidade-de-Dados-Fundamentos-Tecnicas-e-Aplicacoes.pdf. Acesso em: 05 de jul. 2021.

CARLOTO, Selma. **Lei geral de proteção de dados:** enfoque nas relações de trabalho. 1. ed. São Paulo: LTr Editora Ltda, 2020.

CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução a mineração de dados:** conceitos básicos, algoritmos e aplicações. 1. ed. São Paulo. Editora Saraiva, 2016.

CRUTZEN, Rik; PETERS, Gjalt-Jorn Ygram; MONDSCHEIN, Christopher. Why and how we should care about the General Data Protection Regulation. **Routledge Taylor & Francis Group**, Holanda, vol. 34, 2019. Diponível em: https://www.tandfonline.com/doi/pdf/10.1080/08870446.2019.1606222. Acesso em: 09 de agos. 2021.

CYBER attacks on industrial networks and critical infraestructure in brazil grow 860%. In: TI safe. [S.I.], 10 mai. 2021. Disponível em: https://tisafe.com/en/blog/ataques-ciberneticos-em-redes-industriais-e-infraestruturas-criticas-no-brasil-crescem-860. Acesso em: 20 mai. 2021.

DIGITAL, Direito. 5 casos de vazamento de dados nas grandes empresas. In: ASSIS E MENDES. [São Paulo], 19 julho 2018. Disponível em:

https://assisemendes.com.br/vazamento-de-dados-nas-empresas. Acesso em: 20 mai. 2021.

DONDA, Daniel. **Guia prático de implementação da LGPD:** tudo o que sua empresa precisa saber para estar em conformidade. 1. ed. São Paulo: Labrador, 2020.

EMAM, El Khaled; DANKAR, Kamal Fida. Protecting privacy using k-anonymity. **Journal of the American Medical Informatics Association**. v. 15, n. 5, p.627-637, out. 2008. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/pdf/627.S106750270800104 7.main.pdf. Acesso em: 25 jul. 2021.

FONTES, Edison Luiz Gonçalves. **Segurança da Informação:** Gestão e Governança. 1. ed. São Paulo: 2020.

GABINETE PARA A PROTEÇÃO DE DADOS PESSOAIS (GPDP). **Guia para técnicas básicas de anonimização de dados**. [*S.l.*]. abr. 2019. Disponível em: https://www.gpdp.gov.mo/uploadfile/2019/0417/20190417033911965.pdf. Acesso em: 20 jul. 2021.

GARCIA, Lara Rocha. **Lei geral de proteção de dados pessoais (LGPD):** guia de implementação. 1. ed. São Paulo: Blucher, 2020.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6.ed. São Paulo: Atlas, 2018.

GONÇALVES, Eduardo Corrêa. Extração de árvores de decisão com a ferramenta de data mining weka. **IBGE - Instituto Brasileiro de Geografia e Estatística**, [Rio de Janeiro], 2007. Disponível em: https://www.devmedia.com.br/extracao-de-arvores-de-decisao-com-a-ferramenta-de-data-mining-weka/3388. Acesso em: 20 agos. 2021.

GRUPO DE TRABALHO DE PROTEÇÃO DE DADOS DO ARTIGO 29º (GT). **Parecer 05/2014 sobre técnicas de anonimização**. [Macau]. abr. 2014. v.1. Disponível em: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_pt.pdf. Tradução de: Guide to basic data anonymisation techniques. Acesso em: 20 jul. 2021.

K-ANONIMATO: uma introdução. In: PRIVITAR. 7 abr. 2017. Disponível em: https://www.privitar.com/blog/k-anonymity-an-introduction/. Acesso em: 25 jun. 2021.

LI, Ninghui; LI, Tianchenng; Suresh, Venkatasubramanian. t-Closeness: privacy beyond k-anonymity and l-diversity. **Purdue University**, West Lafayette, 22 nov. 2006. Disponível em: https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf. Acesso em: 18 jul. 2021.

LOPES, Everton. Os desafios da lgpd diante da não conformidade corporativa. **Revista Consultor Jurídico**, 29 set. 2020. Disponível em:

https://www.conjur.com.br/2020-set-29/everton-lopes-lgpd-nao-conformidade-corporativa. Acesso em: 22 mar. 2021.

LUBARSKY, Boris. Re-identification of anonymized data. **Georgetown Law Technology Review**. n. 202, p.202-212, 2017. Disponível em: https://georgetownlawtechreview.org/wp-content/uploads/2017/04/Lubarsky-1-GEO.-L.-TECH.-REV.-202.pdf. Acesso em: 20 agos. 2021.

MACIEL, Rafael Fernandes. **Manual Prático sobre a Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/18)**. 1. ed. Goiânia-GO: RM Digital Education, 2019.

MANUAL do aluno. In: PONTIFÍCIA universidade católica de goiás (PUC-GO). Goiás, [2021?]. Disponível em: http://www2.pucgoias.edu.br/manual/. Acesso em: 15 agos. 2021.

MARI, Angelica. Brazil Tech Round: Government Responds To Massive Data Leak, Totvs Results, Healthtech Growth. **Revista Forbs**. 13 fev. 2021. Disponível em: https://www.forbes.com/sites/angelicamarideoliveira/2021/02/13/brazil-tech-round-up-government-responds-to-massive-data-leak-totvs-results-healthtech-growth/?sh=186eef085431. Acesso em: 18 agos. 2021.

MARIN, Maikon Aloan; LOPES, Fabrício Martins. **Indução de árvores de decisão para inferência de redes gênicas**. 2013. Relatório de Pesquisa do Programa de Iniciação Científica, Universidade Tecnológica Federal do Paraná, 2013. Diponível em:http://paginapessoal.utfpr.edu.br/fabricio/fabricio-martins-lopes/pesquisa/orientacoes/relatorio-pibic-2013-maikon-marin.pdf. Acesso em: 24 out. 2021.

MÜLLER, João Artur. O que é a Lei Geral de Proteção de Dados (LGPD) e qual o seu impacto na atividade empresarial online e offline?. In: Martine, Medeiros e Tonetto, 2021. Disponível em: https://www.mmtadvogados.com.br/publicacoes/o-que-e-a-lei-geral-de-protecao-de-dados-lgpd-e-qual-o-seu-impacto-na-atividade-empresarial-online-e-offline. Acesso em: 19 de maio de 2021.

OBJETIVO e abrangência da lgpd**.** In: SERPRO. [S.I., 2018?]. Disponível em: < https://www.serpro.gov.br/lgpd/menu/tratamento-dos-dados/objetivo-e-abrangencia-da-lgpd>. Acesso em: 20 de maio de 2021.

PALUDETTO, Vitor; BARBIERI, Henrique Shirassu. **Guia sobre a Nova Lei Geral de Proteção de Dados Pessoais.** 2019: Edição Kindle.

PANEK, Lin Cristina Tung. **Lei Geral de Proteção de Dados Nº 13.709/2018**: Uma análise dos principais aspectos e do conceito privacidade na sociedade informacional. 2019. Trabalho de Conclusão de Curso (Graduação em Direito, setor de Ciências Jurídicas) – Universidade Federal do Paraná – Faculdade de Direito – UFPR, CURITIBA/PR, 2019. Disponível em: < https://acervodigital.ufpr.br/bitstream/handle/1884/68114/TCC%20FINAL%20-%20lgpd.pdf?sequence=1&isAllowed=y>. Acesso em: 20 jul. 2021.

PECK, Patricia. **Proteção de Dados Pessoais :** comentários à Lei n. 13.709/2018. 2. ed. São Paulo: Saraiva Educação, 2020.

PORTO, Viviane de Araújo. **Descomplicando a Lei Geral de Proteção de Dados Pessoais**. 1.ed. Goiânia-GO, 2020.

PRASSER, Fabian; EICHER, Johanna; SPENGLER, Helmut; BILD, Raffael; KUHN, Klaus A. Flexible data anonymization using ARX – Current status and challenges ahead. **Wiley Online Library**, Berlim, jan. 2020. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1002/spe.2812. Acesso em: 05 de agosto de 2021.

PRASSER, Fabian; KOHLMAYER, Florian. **Putting Statistical Disclosure Control into Practice:** the ARX data anonymization tool. 1. ed. Suiça: Springer, 2015.

QUINLAN, J. Ross. **C4.5**: Programs for machine learning. San Mateo, 1993.

RAVAL, Kalyani M. Data mining techniques, International journal of advanced research in computer science and software enginnering, **Journal of advanced research**, Bhavnagar, GJ, v. 2, ed. 10, p. 439-442, 2012. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.898.6657&rep=rep1&type=pdf. Acesso em: 07 agos. 2021.

SARAVANAN, N; GAYATHRI, V. Performance and classification evaluation of j48 algorithm and kendall's based j48 algorithm (knj48). **International journal of computational intelligence and informatics**, v. 7, n. 4, mar. 2018. Disponível em: https://www.periyaruniversity.ac.in/ijcii/issue/marnew/2_mar_18.pdf. Acesso em: 15 set. 2021.

SILVA, Hebert de Oliveira. **Uma Abordagem Baseada em Anonimização para Privacidade de Dados em Plataformas Analíticas**. 2019. Dissertação (Mestrado em Tecnologia, área de Sistemas de Informação e Comunicação) – Universidade Estadual de Campinas – Faculdade de Tecnologia – FT/UNICAMP, LIMEIRA/SP, 2019. Disponível em: http://repositorio.unicamp.br/jspui/bitstream/REPOSIP/334676/1/Silva_HebertDeOliveira_M.pdf. Acesso em: 20 jul. 2021.

SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados**: com aplicações em R. 1. ed1. Rio de Janeiro: Elsevier, 2016.

SOARES, Paulo Vinicius de Carvalho. GUIA LGPD: Lei Geral de Proteção de Dados SIMPLIFICADA. LGPDBRASIL, 14 de agosto de 2019. Disponível em: <https://conteudo.lbca.com.br/lgpd-guia-simplificado?fbclid=IwAR1wjAQFB8IKtw6AZ1PsoFsLRwSgAWpDN93narXK7FM_VW3lRrWDFNitR1g >. Acesso em: 9 de maio de 2021.

SUNDARESAN, Sumathie. Thesis, 2016. 25 slides. Disponível em: https://slideplayer.com/slide/9180203/. Acesso em: 10 jul. 2021.

TERROVITIS, Manolis. **D9.6 DATA ANONYMIZATION SERVICES**. 0.3 v. Ares, 2017. Disponível em: https://www.openaire.eu/d9-6-data-anonymization-services/view-document. Acesso em: 25 agos. 2021.

WAZLAWICK, Raul Sidnei. **Metodologia de pesquisa para ciência da computação**. 2. ed. Rio de Janeiro: Elsevier, 2014.

WITTEN, Ian H. et al. **The weka workbench.** data mining practical machine learning tools and techniques. 4. ed. Morgan Kaufmann, 2016. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf. Acesso em: 21 agos. 2021.

ANEXO I – Termo de publicação de produção acadêmica.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PRÓ-REITORIA DE GRADUAÇÃO

Av. Universitária, 1069 ● Setor Universitário
Caixa Postal 86 ● CEP 74605-010
Goiânia ● Goiás ● Brasil
Fone: (62) 3946.1021 I Fax: (62) 3946.1397
www.pucgoias.edu.br I prograd@pucgoias.edu.br

RESOLUÇÃO 038/2020 - CEPE
ANEXO I

Termo de autorização de publicação de produção acadêmica

O estudante Felipe Silva Paula do Curso de Ciência da Computação, matrícula 2016.1.0028.0086-9, telefone: 62 985933846, e-mail fellype-tecnovation@outlook.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponbilizar o Trabalho de Conclusão de Curso intitulado 'Efeitos da anonimização nos processos de mineração de dados' gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF), específicos da área para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, __15_ de dezembro de 2021

Assinatura do autor: Felipe Silva Paula

Nome completo do autor: Felipe Silva Paula

Assinatura do professor – orientador: Sibelius Lellis Vieira

Nome completo do professor – orientador: Sibelius Lellis Vieira