UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Proteção de Dados Pessoais na Era da Inteligência Artificial

Felipe Casali Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Felipe Casali Silva

Proteção de Dados Pessoais na Era da Inteligência Artificial

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Cristina Dutra de Aguiar

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Casali Silva, Felipe
C334p Proteção de Dados Sensíveis na Era da Inteligência Artificial / Felipe Casali Silva; orientadora Profa. Dra. Cristina Dutra de Aguiar . -- São Carlos, 2023.
69 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Inteligência Artificial. 2. Proteção de Dados. 3. LGPD. 4. Compliance. 5. Big Data. I. , Profa. Dra. Cristina Dutra de Aguiar, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Felipe Casali Silva

Personal Data Protection in the Age of Artificial Intelligence

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Profa. Dra. Cristina Dutra de Aguiar

Original version

São Carlos

2023

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

Dedico este trabalho à minha primeira e maior incentivadora nos estudos na área de TI, minha falecida avó Iria Malicia Casali, que já em meados de 95 insistia em me dizer que deveria trilhar o caminho da informática pois este seria o futuro!

Também agradeço aos meus pais que sempre incentivaram e apoiaram o caminho do aprendizado e do estudo, fazendo grandes esforços para proporcionar um ensino de qualidade dentro de suas possibilidades.

Por fim, deixo registrada minha dedicatória a todos que contribuíram com ideias para o desenvolvimento deste trabalho, incluindo aqui os professores do ICMC da USP que proporcionaram uma incrível imersão no mundo da IA durante a jornada do MBA.

AGRADECIMENTOS

Em primeiro lugar, agradeço a minha esposa Thatiane Cabral Casali, por contribuir imensamente para que essa jornada do MBA em Inteligência Artificial e BigData pudesse acontecer. Sem ela não teria conseguido!

Não poderia deixar de mencionar meus filhos, Arthur Cabral Casali (9) e Olivia Cabral Casali (5), respeitaram as horas de dedicação necessárias ao desenvolvimento deste trabalho. Obrigado meus filhos!

Aos professores e orientadores do ICMC da USP, meu muito obrigado por compartilhar conhecimento, e permitir que o tema da Inteligência Artificial possa ser difundido e aprimorado em toda a comunidade científica.

Por fim, agradeço a todos os colegas de trabalho, amigos e pessoas de meu convívio que me incentivaram e apoiaram nessa grande jornada.

"Quando penso que cheguei ao meu limite, descubro que tenho forças para ir além."
Ayrton Senna

RESUMO

SILVA, F.C. **Proteção de Dados Pessoais na Era da Inteligência Artificial**. 2023. 74p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A Inteligência Artificial (IA) trouxe avanços importantes para diversos setores econômicos nos últimos anos. Pode-se destacar a avalanche de dados capturados e armazenados, para posteriormente serem agregados e processados por algoritmos capazes de gerar modelos avançados para predição de eventos. Como resultado, novas oportunidades surgiram, sendo que as empresas vislumbraram a possibilidade de realizar a tomada de decisões com mais assertividade, direcionar investimentos e até otimizar processos pré-existentes. Contudo, a manipulação massiva de dados introduz diversos desafios. Dentre esses desafios, destaca-se a necessidade de controlar o uso dos dados sob a perspectiva de proteger a privacidade dos indivíduos. Neste sentido, diversos países, grupos econômicos e entidades reguladoras têm criado mecanismos para estabelecer regras quanto ao uso de dados pessoais, que são dados sensíveis. Dentro deste contexto, este trabalho de conclusão de curso tem como objetivo investigar o tema proteção de dados na era da IA. São abordados desafios e técnicas de proteção de dados que permitam o avanço da IA respeitando a privacidade dos indivíduos, de forma que os negócios e a sociedade em geral possam se beneficiar desses avanços. A investigação é realizada considerando aspectos teóricos e práticos. Do ponto de vista teórico, é feito um estudo de exemplos de fontes de dados contendo dados sensíveis, é realizado um delineamento do clido de vida dos dados, são sumarizados os principais aspectos relacionados às leis e regulamentações para proteção dos dados, são identificadas e discutidas técnicas e ferramentas voltadas à proteção dos dados e são definidas sete estratégias que um cientista de dados deve empregar para ajudar na proteção dos dados em projetos de IA. Do ponto de vista prático, é desenvolvido um modelo de IA baseado em Regressão Logística, TF-IDF e expressões regulares, o qual é aplicado sobre dados sensíveis da LGPD.

Palavras-chave: Inteligência Artificial. Privacidade. LGPD. Dados Pessoais. Dados Sensíveis.

ABSTRACT

SILVA, F.C. **Personal Data Protection in the Age of Artificial Intelligence**. 2023. 74p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Artificial Intelligence (AI) registered unprecedent advances to several economic sectors in recent years. The amount of data captured and stored, to later be aggregated and processed by algorithms capable of generating advanced models for predicting events, is increasing every day. As a result, new opportunities have emerged, and companies are working with the possibility of making more precise business decisions, directing investments and even optimizing pre-existing processes. However, massive data manipulation introduces several challenges and risks. Among these challenges, the need to control the use of data from the perspective of protecting the privacy of individuals stands out. In this sense, several countries, economic groups and regulatory entities have created mechanisms to establish rules regarding the use of personal data, which are sensitive data. Within this context, this course completion work aims to investigate the subject of data protection in the AI era. Challenges and data protection techniques that allow the advancement of AI respecting the privacy of individuals are addressed, so that business and society in general can benefit from these advances. The investigation is composed by theoretical and practical aspects. From a theoretical point of view, a study is made of examples of data sources containing sensitive data, an outline of data life cycle in AI projects, the main aspects related to laws and regulations for data protection are summarized, techniques and tools for data protection are discussed and seven strategies that a data scientist should employ to help protect data in AI projects are defined. From a practical point of view, an AI model based on Logistic Regression, TF-IDF and regular expressions is developed, which is applied to find sensitive LGPD data.

Keywords: Artificial Intelligence. LGPD. Governance. Personal Data. Sensitive Data. Data Masking. Data Anonimization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IA	Inteligência Artificial
LGPD	Lei Geral de Proteção de Dados
GDPR	General Data Protection Rules
HIPAA	Health Insurance Portability and Accountability Act
DPO	Data Protection Officer
ABNT	Associação Brasileira de Normas Técnicas
ETL	Extract, Transfer, Load
IBGE	Instituto Brasileiro de Geografia e Estatística
DW	Data Warehouse
DL	Data Lake
USPSC	Campus USP de São Carlos
SGDB	Sistemas de Gerenciamento de Banco de Dados
SGA	Sistemas de Gerenciamento de Arquivos

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo é feita a introdução deste trabalho de conclusão de curso, cujo objetivo consiste em investigar o tema proteção de dados na era da Inteligência Artificial (IA). A investigação é realizada considerando aspectos teóricos e práticos. Os aspectos teóricos visam correlacionar dois assuntos pouco explorados em artigos acadêmicos das áreas de exatas, mas muito discutidos em nossa sociedade nos últimos anos: IA e *Compliance*. Os aspectos práticos visam aplicar modelos de IA para treinar dados sensíveis. O capítulo está estruturado da seguinte forma. Na seção 1.1 é feita a contextualização do trabalho. Nas seções 1.2 e 1.3 são discutidos aspectos relacionados às regulamentações e ao vazamento dos dados. Na seção 1.4 são descritas as justificativas e as motivações para o desenvolvimento do trabalho. Na seção 1.5 são listados os objetivos e as contribuições. O capítulo é finalizado na seção 1.6, na qual é descrita a estruturação da monografia.

1.1 Contextualização

As discussões sobre a proteção de dados pessoais compõem a pauta de governos e empresas em todo o mundo há alguns anos. Regulamentações como LGPD (Lei Geral de Proteção de Dados), GDPR (General Data Protection Regulation) e HIIPA (Health Insurance Portability and Accountability Act), dentre outras, definem regras claras sobre o uso de dados pessoais. Contudo, essas regulamentações não possuem tratativas específicas sobre o uso dos dados em projetos de IA.

Nos últimos meses, a evolução acelerada da IA, e principalmente das técnicas de PLN (Processamento de Linguagem Natural), têm imposto um novo panorama. Neste panorama, surgem dúvidas e desconfianças sobre como os dados estão sendo protegidos em projetos que envolvem *Big Data* e IA. Essas dúvidas têm gerado desconfiança e medo. Adicionalmente, elas motivaram a criação de grupos específicos para tratar sobre o tema.

De certa forma, alguns desses grupos têm discutido formas de proteger os dados. Contudo, a falta de entendimento tem gerado um cenário no qual se cogita regular o uso da IA, definindo onde esse tipo de tecnologia pode ou não ser aplicado. Entretanto, o caminho da regulação por meio da proibição de uso em alguns setores pode ser um grande retrocesso. Dependendo da forma como grandes grupos decidirem tratar esse tema, há uma grande probabilidade de haver uma queda nos benefícios que a tecnologia pode trazer para a sociedade.

1.2 Regulamentações

Existem inúmeras discussões em andamento, em várias camadas da sociedade, que visam definir regras, formas de controle e até penalidades para quem desrespeita as regulamentações de proteção de dados. A maioria dos trabalhos acadêmicos que discorrem sobre proteção de dados têm como origem cursos de Direito.

Em seu trabalho de doutorado pela Faculdade de Direito da USP, Renata Capriolli, Queiroz (2023), investiga diversas nuances do ponto de vista jurídico sobre as responsabilidades de um novo profissional. Esse, chamado de **encarregado de dados**, tem como objetivo atuar como um canal de comunicação entre instituição, os titulares dos dados e a Autoridade Nacional de Proteção de Dados[1] (ANPD).

Capriolli também aborda detalhes sobre regulamentações como LGPD e GDPR (doravante RGDP, na sigla em português, Regulamento Europeu de Proteção de Dados). Essas regulamentações têm como objetivo proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural. Elas estabelecem, de forma clara, que existem regras a serem respeitadas no tocante ao uso de dados pessoais para quaisquer que sejam as finalidades.

Apesar de todas essas iniciativas e uma quantidade significativa de literatura abordando o tema da proteção de dados, é muito comum as empresas não saberem como lidar com os dados sensíveis em sua forma mais básica: o próprio armazenamento. As dúvidas aumentam à medida que os dados são considerados fonte de informação para *pipelines* de dados, obtenção de *insights* e projetos que têm como base a manipulação e o processamento de diferentes fontes de dados.

Nas grandes empresas, contrapõem-se duas perspectivas de interesse. Por um lado, posicionam-se os DPOs (Data Protection Officer), os quais são responsáveis por definir regras de controle no acesso e uso dos dados. Por outro lado, encontram-se os times de negócios e os cientistas de dados, que precisam ter acesso aos dados que serão base para a tomada de decisões estratégicas. Atender a essas duas perspectivas de interesse é de suma importância no que tange à IA e o *compliance*.

1.3 Vazamentos de Dados

Outro ponto importante que tem influência direta sobre a gestão dos dados é a prevenção de vazamentos. Nos últimos anos, os dados deixaram de ser um legado, e passaram a ter um valor incalculável para as empresas. Como resultado, surgiram novos desafios

[1] BRASIL. Presidência da República, Autoridade Nacional de Proteção de Dados. **Portaria n.o 11, de 27 de janeiro de 2021**. Torna pública a agenda regulatória para o biênio 2021-2022. Brasília, DF: Presidência da República, 2021a. Disponível em: https://www.in.gov.br/en/web/dou/-/portaria-n-11-de-27-de-janeiro-de-2021-301143313. Acesso em: 07 jun. 2023.

relacionados à prevenção de ataques provindos de organizações criminosas. Mecanismos sofisticados e automáticos de ataque, criados por grupos de *hackers* espalhados por todo o mundo, trabalham em regime 24x7 na tentativa de sequestrar dados das empresas. O objetivo consiste em obter ganhos ilícitos com pedidos de resgate. Adicionalmente, os *hackers* também realizam chantagens para não expôr dados sensíveis na mídia ou em redes de conteúdo não indexadas como a Deep Web [2] ou Dark Web [3].

O vazamento de dados geralmente vem acompanhado de grandes prejuízos, sendo alguns deles tangíveis, e outros não. Como forma de prejuízo tangível, pode-se citar multas e penalidades estabelecidas nas regulamentações já mencionadas. Contudo, se torna impossível calcular o tamanho do impacto na imagem dos negócios quando um vazamento de dados se torna público, ou mesmo quando esses dados vazados são disponibilizados em redes clandestinas.

Também pode-se citar resgates com valores exorbitantes, os quais fazem parte das estratégias dos *hackers* que sequestram os dados em ataques do tipo *ransomware*. No artigo Evolution of ransomware, os autores (O'KANE; SEZER; CARLIN, 2018) detalham todas as características desse tipo de ataque, que tem como principal objetivo o sequestro de dados.

> "Ocorre, assim, o uso da tecnologia para fins sombrios, corrompendo-se relações sociais, com o intuito de chantagear, obter vantagens econômicas ou causar constrangimentos na condição de conseguir algum favor ou beneficiamento indevido, sempre pelo meio da força de suas ações. A questão desses crimes virtuais se tornou rapidamente uma séria realidade, levando os operadores e estudiosos do Direito a empreenderem maiores esforços para abarcar a sua complexidade, que configuraram situações antes sequer consideradas relevantes." (FORNASIER; SPINATO; RIBEIRO, 2020)

Algumas vezes o resgate é calculado com base na quantidade de registros extraviados. Com isso, já são vários os relatos de empresas e governos que efetuaram a transferência de altos valores na tentativa de minimizar os impactos do vazamento de dados. Vale ressaltar que esse tipo de pagamento não oferece nenhuma garantia de que os dados extraviados serão devolvidos ou eliminados de forma definitiva.

Arquiteturas de rede compostas por ambientes locais (*on-premisses*) e nuvem (*cloud*) aumentam o tamanho do desafio pois os dados esparsos precisam ser constantemente movimentados e agregados por processos de ETL, e sem a devidas precauções, os riscos de vazamento se tornam mais latentes.

[2] https://en.wikipedia.org/wiki/Deep_web
[3] https://en.wikipedia.org/wiki/Dark_web

1.4 Justificativa e Motivação

Em geral, toda tecnologia disruptiva vem acompanhada de grandes desafios. Com a IA isso não é diferente, e um dos principais pontos de discussão é a privacidade. Com dezenas de petabytes sendo capturados e processados, é natural que muitos dados sejam referentes a informações pessoais e/ou sensíveis, e até mesmo sigilosas, que precisam ter um tratamento adequado.

Ao associar grandes volumes de dados às tecnologias modernas como IA, são criadas novas possibilidades. Por exemplo, a alta capacidade computacional associada aos algoritmos avançados pode revelar comportamentos e características de indivíduos com base em registros de informações provenientes de diferentes origens. Um exemplo consiste na utilização de algoritmos que podem apontar preferências pessoais como religiosas e sexuais sobre os usuários por meio do cruzamento de dados de GPS capturados constantemente por *smartphones* e *smartwatches* com os tipos de estabelecimentos mais frequentemente visitados por esses usuários.

Em outro exemplo, se uma pessoa frequentemente visita uma igreja católica, um algoritmo de IA treinado com base em dados sobre os tipos de estabelecimentos existentes em determinada área pode identificar a preferência religiosa desta pessoa. Contudo, essa informação, considerada pessoal, deve permanecer em sigilo, há menos que o titular dos dados autorize expressamente que essa informação pode ser divulgada em meios públicos.

Antes de implementar qualquer projeto de proteção de dados, é preciso entender quais são as diferenças entre **dados pessoais e dados sensíveis**. Além disso, é primordial entender como esses tipos de dados se relacionam, para garantir que a agregação de dados de diferentes fontes não represente um risco de exposição de dados indevida. Também é preciso abordar temas que permeiam a IA, pois todas as etapas que envolvem um projeto de IA, desde a coleta do dado bruto até o conhecimento gerado e utilizado para a identificação de *insights*, requerem que a privacidade seja respeitada.

É importante destacar que existem evidências de que grande parte dos vazamentos de dados não ocorrem em ambientes de produção, e sim em ambientes não-produtivos, dentre eles ambientes de *Analytics*.

> "A definição do termo *Analytics* abrande uma infinidade de conceitos relacionados ao estudo de sistemas de suporte à decisão. Sob a perspectiva de gestão organizacional, pode ser compreendida como uma série de procedimentos com o extensivo uso de dados, análises estatísticas e quantitativas, modelos preditivos ou descritivos, resultando na geração de informações de valor para o negócio e apoiando a geração de vantagem competitiva mediante o suporte à rápida tomada de decisões" (SOUSA, 2018)

Dada a importância que o tema da proteção de dados tem na sociedade, o principal motivador desse trabalho consiste em apresentar alternativas, técnicas e possibilidades futuras para que a IA possa continuar evoluindo e melhorando a vida das pessoas sem colocar em risco a privacidade.

1.5 Objetivos

Desmistificar os riscos associados ao uso de dados em projetos de IA consiste no principal objetivo desse trabalho. Para tanto, são abordados desafios e técnicas de proteção de dados que permitam o avanço da IA respeitando a privacidade dos indivíduos, de forma que os negócios e a sociedade em geral possam se beneficiar desses avanços.

A investigação é realizada considerando aspectos teóricos e práticos. Os aspectos teóricos visam correlacionar dois assuntos pouco explorados em artigos acadêmicos das áreas de exatas, mas muito discutidos em nossa sociedade nos últimos anos: IA e *Compliance*. Neste sentido, é feito um estudo teórico completo que consiste na análise do caminho dos dados desde as suas fontes primárias, até o seu uso em algoritmos de IA. São mapeados os riscos associados. Também são estabelecidas boas práticas para que esses riscos sejam mitigados durante o desenho da arquitetura do projeto.

Os aspectos práticos visam aplicar modelos de IA para treinar dados sensíveis. Neste sentido, é utilizado um modelo baseado em regressão logística para mostrar como dados sensíveis podem ser mapeados com precisão e velocidade, podendo ser utilizados no desenvolvimento de projetos de IA.

1.6 Estruturação da Monografia

Além deste capítulo introdutório, esta monografia é composta pelos seguintes capítulos:

- Capítulo 2. Descreve diferentes exemplos de fontes de dados com o objetivo de destacar que essas fontes possuem dados sensíveis e que devem ser manipulados de forma apropriada para garantir *compliance*.
- Capítulo 3. Descreve aspectos relacionados ao caminho dos dados desde as suas fontes primárias até o uso em algoritmos de IA.
- Capítulo 4. Detalha aspectos relacionados às leis e às regulamentações para proteção dos dados.
- Capítulo 5. Discute os resultados do aspecto teórico investigado neste trabalho, os quais são relacionados a como identificar e proteger dados sensíveis.

- Capítulo 6. Detalha os resultados do aspecto prático investigado neste trabalho, os quais são relacionados ao uso de regressão logística na manipulação de dados sensíveis.
- Capítulo 7. Conclui a monografia.

2 NEGÓCIOS: AS GRANDES FONTES DE DADOS

Neste capítulo são descritos diferentes exemplos de fontes de dados com o objetivo de destacar que essas fontes possuem dados sensíveis e que devem ser manipulados de forma apropriada para garantir *compliance*. Antes de descrever esses exemplos, na seção 2.1 é delineado um panorama geral a respeito da avalanche de dados produzida atualmente, a qual é possível devido aos avanços em IoT (Internet of Things) e tecnologia 5G. Na sequência, são detalhados as seguintes fontes de dados: redes sociais (seção 2.2), comércio eletrônico (seção 2.3), bancos digitais (seção 2.4), saúde e outros setores econômicos (seção 2.5).

2.1 Avalanche de Dados, IoT e o 5G

Os últimos 20 anos foram marcados por diversas novidades tecnológicas que se popularizaram e hoje são grandes fontes de geração de dados. Um exemplo é a evolução dos aparelhos celulares. Há 20 anos atrás, esses aparelhos serviam apenas para fazer ligações telefônicas e tinham um custo bem elevado, o que limitava bastante o acesso da população em geral a essa tecnologia.

Contudo, os aparelhos celulares foram se popularizando. Atualmente, eles possuem diversos recursos, como sensores, câmeras e aplicativos como redes sociais. Como resultado, eles são responsáveis por boa parte do tráfego de dados na internet em nível global[1].

De acordo com a empresa de pesquisa Statista, em 2022 o número de usuários ativos de *smartphones* em todo o mundo ultrapassou a quantidade de 5 bilhões de usuários, ou seja, mais de 60% da população mundial naquele ano[2]. Para efeito de comparação, a empresa de pesquisas Hedges & Company publicou um relatório[3] em 2018 apontando que existiam na época aproximadamente 1.4 bilhão de motoristas de veículos ativos em todo o mundo. Em números absolutos, pode-se dizer que o número de celulares em 2022 era 3 vezes maior do que o número de carros em todo o planeta.

Ainda de acordo com a Statista, o celular é responsável por aproximadamente metade do tráfego da web em todo o mundo. No primeiro trimestre de 2023, os dispositivos móveis (excluindo tablets) geraram 58,33% do tráfego global, oscilando consistentemente em torno da marca de 50% desde o início de 2017, antes de ultrapassá-la permanentemente

[1] https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/
[2] https://www.statista.com/topics/779/mobile-internet/#topicOverview
[3] hedgescompany.com/blog/2018/10/number-of-licensed-drivers-usa/

em 2020[4]. Em 2020, o volume de dados trafegados de aparelhos celulares[5] estava estimado em 36 exabytes por mês, ou seja, em um ano 412 exabytes de informações preciosas sendo geradas por bilhões de usuários em todo planeta.

Além dos aparelhos celulares, outros tipos de dispositivos inteligentes, também conhecidos como *Gadgets*, tiveram um crescimento exponencial. Um exemplo clássico são os *smartwatches*, ou relógios inteligentes. Eles são equipados com diversos tipos de sensores, que hoje são capazes de identificar problemas de saúde, acidentes ou quedas dos usuários. Em um futuro não muito distante, os relógios inteligentes serão capazes de gerar alertas com antecedência sobre sintomas iniciais de doenças e recomendar tratamentos, ou até mesmo acionar serviços de emergência informando a localização do usuário em tempo real.

Todos esses dispositivos móveis são considerados parte do IoT. Esse conceito foi criado em 1999 por Kevin Ashton (ASHTON *et al.*, 2009), um pioneiro da tecnologia britânico. Na época, Ashton trabalhava para a empresa de tecnologia Procter & Gamble e estava pesquisando formas de melhorar a eficiência do gerenciamento de estoque usando a tecnologia RFID (Identificação por Radiofrequência). Enquanto trabalhava nesse projeto, Ashton percebeu que a tecnologia RFID poderia ser usada para conectar objetos do mundo real à internet, criando uma rede de dispositivos interconectados capazes de coletar e compartilhar dados. Foi nesse contexto que Ashton cunhou o termo “Internet das Coisas” .

IoT representa um dos principais desafios quando o tema é a privacidade dos usuários, mesmo antes de haver qualquer associação com a IA. A quantidade, variedade e velocidade dos dados gerados por esses dispositivos é impressionante. Na ótica de segurança dos dados, a preocupação é grande pois hoje existem algumas centenas de fabricantes desses dispositivos investindo milhões de dólares para incrementar funcionalidades e recursos que visam manter a fidelidade de seus clientes, e muitas dessas soluções usam dados.

Por conta de todo esse contexto de evolução tecnológica e da produção de dados dentro do conceito de *big data*, é comum que organizações focadas em privacidade dos usuários efetuem questionamentos constantes sobre como esses dados são utilizados, e quais são os riscos para a sociedade.

2.2 Redes Sociais

As redes sociais são plataformas digitais que permitem que indivíduos e organizações compartilhem informações, ideias e interesses em um ambiente virtual. Com o passar do tempo, as redes sociais se tornaram uma fonte valiosa de dados, principalmente pela facilidade com que as preferências dos usuários podem ser mapeadas. Essas preferências

[4] https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/

[5] https://www.statista.com/statistics/270878/global-consumer-ip-data-volume-by-connection-type/

oferecem suporte para a geração de *insights* comerciais valiosos para empresas privadas, além de servir de fonte de dados para órgãos de governo e segurança como o FBI.

Por exemplo, o FBI utiliza o recurso SOMEX (social media exploitation)[6] para facilitar a identificação de responsáveis por ameaças de ordem pública ou à vida. Todas as postagens, fotos, conversas nos *chats* e demais registros nas redes sociais podem estar disponíveis para que agentes de segurança inspecionem e investiguem, a fim de garantir a ordem e a segurança.

Alguns casos recentes de vazamentos de dados em redes sociais demonstram claramente que existe um valor elevado nos dados de seus usuários. Esses vazamentos impactaram milhões de usuários em todo o mundo. Os maiores deles são:

- **Facebook:** Em 2018, o Facebook foi atingido por um grande escândalo de privacidade envolvendo a empresa de consultoria política Cambridge Analytica. A Cambridge Analytica usou informações pessoais de cerca de 87 milhões de usuários do Facebook sem consentimento para criar anúncios políticos altamente personalizados[7].

- **LinkedIn:** Em junho de 2021, foi relatado que um arquivo de dados contendo informações pessoais de 700 milhões de usuários do LinkedIn estava sendo vendido na dark web. As informações incluíam nomes completos, endereços de e-mail, números de telefone e outras informações de perfil[8].

- **Twitter:** Em maio de 2020, foi relatado que dados pessoais de cerca de 330 milhões de usuários do Twitter foram expostos na dark web. Os dados incluíam nomes de usuário, endereços de e-mail e senhas criptografadas[9].

- **Facebook:** Em novembro de 2022, foi relatado que dados de 1,5 bilhões de usuários do aplicativo Facebook estavam sendo vendidos em um fórum Hacker. As informações incluíam nomes de usuários, números de telefone, endereços de e-mail e outras informações pessoais[10].

Esses vazamentos reforçam a tese de que os dados de redes sociais têm um valor inestimável, e são cada vez mais cobiçados por organizações criminosas online. Considerando o tema principal deste trabalho de conclusão de curso, destaca-se que em muitos projetos de IA os dados das redes sociais são usados para analisar comportamentos, tendencias e

6 https://techxplore.com/news/2022-09-fbi-agents-social-media-domestic.html
7 https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought
8 https://restoreprivacy.com/linkedin-data-leak-700-million-users/
9 https://www.securityreport.com.br/overview/pesquisa-indica-como-dados-vazados-do-twitter-podem-ser-vendidos-na-dark-web/
10 https://www.privacyaffairs.com/facebook-data-sold-on-hacker-forum/

tomar decisões. Portanto, é necessário utilizar apenas os dados necessários para serem processados no contexto de IA, evitando assim que as aplicações desenvolvidas fiquem vulneráveis aos ataques de hackers.

2.3 Comércio Eletrônico

O comércio eletrônico tem crescido significativamente nos últimos anos, impulsionado pela popularização da internet e pelo aumento da confiança dos consumidores em fazer compras online. A pandemia da COVID-19 também impulsionou ainda mais o crescimento do comércio eletrônico, já que muitos consumidores foram forçados a comprar online devido às restrições de quarentena.

De acordo com a Statista, o comércio eletrônico global deve atingir a marca de USD 3.64 trilhões em 2023. Além disto, mantida a taxa de crescimento atual de cerca de 11.16%, em 2027 esse mercado pode atingir um faturamento global de USD 5.56 trilhões[11].

Relatórios da plataforma similarweb [12] apontam que os setores de comércio eletrônico que mais crescem incluem moda, eletrônicos, beleza e cuidados pessoais, e alimentação e bebidas. Muitas empresas tradicionais também estão mudando para o comércio eletrônico para atender à demanda dos consumidores e permanecerem competitivas.

Em resumo, o comércio eletrônico tem experimentado um crescimento significativo nos últimos anos e deve continuar a crescer à medida que mais consumidores mudam para compras online.

O comércio eletrônico representa uma grande fonte de dados para projetos de IA, oferecendo suporte para o estudo e previsão do comportamento dos usuários. Como resultado, pode trazer grandes vantagens competitivas. Porém, o comércio eletrônico representa um cenário repleto de dados sensíveis, com informações pessoais, endereços, preferências de consumo e perfil econômico. Esses dados sensíveis representam um grande risco de *compliance* caso não sejam protegidos da maneira apropriada.

2.4 Bancos Digitais

Não há um número exato de bancos digitais no mundo, pois eles estão surgindo continuamente e as definições de bancos digitais podem variar. No entanto, segundo o relatório do Banco Interamericano de Desenvolvimento (BID) publicado em 2021, existem mais de 3.000 *fintechs* (empresas que utilizam tecnologia para oferecer serviços financeiros) na América Latina, incluindo bancos digitais[13].

[11] https://www.statista.com/outlook/dmo/ecommerce/worldwide
[12] https://www.similarweb.com/
[13] https://publications.iadb.org/publications/portuguese/viewer/Relatorio-anual-2021-do-Banco-Interamericano-de-Desenvolvimento-Resenha-do-ano.pdf

Segundo o relatório FINTECH na América Latina e Caribe publicado pelo Banco Interamericano de Desenvolvimento (BID) em 2022, o ecossistema das *fintechs* na América Latina e no Caribe cresceu 112% desde que foi publicada, em parceria com a Finnovista, a última edição da análise sobre esse setor em 2018. Considerando América Latina e Caribe, a quantidade de *fintechs* passou de 1.166 plataformas para 2.482 em pouco mais de três anos. A concentração do número de plataformas mudou pouco em relação à publicação anterior, e continua sendo liderada pelo Brasil (31% do total), seguido por México (21%), Colômbia (11%), Argentina (11%) e Chile (7%).

Os bancos digitais (neobancos), são instituições financeiras que oferecem serviços bancários 100% digitais, sem a necessidade de agências físicas. Eles surgiram na última década como uma alternativa aos bancos tradicionais e têm ganhado popularidade devido à sua conveniência, taxas mais baixas e experiência do usuário mais moderna.

No entanto, como qualquer instituição que gerencia dados financeiros confidenciais, os bancos digitais estão sujeitos aos riscos de segurança cibernética e vazamentos de dados. Em alguns casos, houve vazamentos de dados significativos de bancos digitais, expondo informações pessoais de clientes.

Por exemplo, em 2022, o Banco Central confirmou o vazamento [14] de 160.147 chaves do Pix que estavam sob responsabilidade do banco digital. Na ocasisão, não houve confirmação se o vazamento também afetou dados sensíveis, como senhas, saldos e histórico de movimentações. Em outra situação semelhante, também em 2022, o Banco Pan confirmou o vazamento de dados de clientes, onde o volume de contas comprometidas, de acordo com denúncias anônimas, chegou à casa dos milhões. [15]

Em nenhum dos exemplos citados os vazamentos estavam ligado diretamente à projetos de IA, contudo, esses acontecimento reforçam a tese de que os bancos digitais possuem dados valiosos e que precisam ser protegidos. Neste contexto, salienta-se novamente que os cuidados com a privacidade dos dados dos usuários são de suma importância para que os projetos não ofereçam riscos de exposição de dados confidenciais e financeiros dos clientes.

2.5 Saúde e outros setores econômicos

O uso de dados no ramo da saúde tem se tornado cada vez mais relevante e impactante nos últimos anos, impulsionado pelo avanço da tecnologia e pela crescente disponibilidade de informações digitais. Essa tendência tem o potencial de transformar a prestação de cuidados de saúde, melhorar a eficiência operacional, impulsionar pesquisas

[14] https://www.cnnbrasil.com.br/economia/banco-central-informa-vazamento-de-dados-do-pix-em-instituicao-financeira/

[15] https://www.cnnbrasil.com.br/economia/banco-central-informa-vazamento-de-dados-do-pix-em-instituicao-financeira/

médicas e, acima de tudo, salvar vidas.

Algumas das principais áreas em que o uso de dados está desempenhando um papel fundamental na área da saúde são:

- **Informações do paciente:** Os sistemas eletrônicos de registro de saúde (EHRs) têm substituído gradualmente os registros de papel, possibilitando o armazenamento e o acesso eficiente às informações do paciente. Esses dados são essenciais para os profissionais de saúde fornecerem um atendimento personalizado, acompanharem o histórico médico e melhorarem a coordenação entre diferentes especialistas.

- **Análise e diagnóstico:** A análise de dados na área da saúde ajuda a identificar padrões e tendências em grandes conjuntos de informações médicas, possibilitando diagnósticos mais precisos e precoces. Algoritmos de aprendizado de máquina podem ser treinados para detectar doenças, interpretar exames médicos (como radiografias e ressonâncias magnéticas) e sugerir tratamentos adequados.

- **Monitoramento de pacientes:** Dispositivos vestíveis (*wearables*) e sensores de saúde conectados permitem o monitoramento contínuo de sinais vitais, atividades físicas e padrões de sono dos pacientes. Esses dados em tempo real podem ser usados para prevenir problemas de saúde, fornecer alertas precoces e melhorar a gestão de doenças crônicas.

- **Medicina de precisão:** Com base nas informações genéticas e moleculares dos pacientes, a medicina de precisão usa dados para personalizar tratamentos com maior eficácia e menores efeitos colaterais. Isso permite o desenvolvimento de terapias direcionadas para grupos específicos de pacientes.

- **Pesquisa médica:** O compartilhamento de dados entre instituições e pesquisadores permite que estudos médicos sejam conduzidos em escalas maiores e com maior representatividade. Isso acelera a descoberta de novos tratamentos, drogas e avanços na área da saúde.

- **Previsão e prevenção:** Com base em dados históricos de saúde da população, é possível usar análises preditivas para identificar grupos de alto risco para determinadas doenças, melhorando o planejamento de intervenções preventivas e políticas de saúde pública.

Apesar dos inúmeros benefícios do uso de dados na saúde, também surgem desafios relacionados à privacidade e segurança das informações do paciente. Portanto, é fundamental garantir a conformidade com as regulamentações de proteção de dados, como o HIPAA (Lei de Portabilidade e Responsabilidade do Seguro de Saúde) nos Estados Unidos

e o RGPD na União Europeia, para garantir a confidencialidade e integridade dos dados do paciente.

Em resumo, o uso de dados na área da saúde apresenta oportunidades significativas para melhorar a qualidade do atendimento, facilitar pesquisas médicas e avançar a medicina para um futuro mais personalizado e eficiente. No entanto, o uso responsável e ético dos dados é essencial para garantir o respeito à privacidade dos pacientes e o sucesso dessa transformação digital na saúde.

3 DADOS: O CICLO DE VIDA, DA FONTE AO ALGORITMO

A IA possui suporte de diversas tecnologias que surgiram nas últimas duas décadas. Essas tecnologias vêm sendo aprimoradas para permitir que os dados sejam utilizados com velocidade e escala. O objetivo é prover suporte para a descoberta de conhecimento e para a solução de problemas complexos.

A tomada de decisão inteligente, que hoje é disciplina obrigatória em todos os setores econômicos, oferece diversos diversos benefícios, dentre os quais pode-se citar: (i) melhoria de processos; (ii) controle de gastos; (iii) maior produtividade; (iv) melhores produtos e soluções; (v) inovação; (vi) melhor tempo de reação à intempéries no mercado; e (vii) direcionamento adequado dos investimentos.

Como pilar fundamental para todos os modelos de IA, o ciclo de vida dos dados precisa ser entendido, mapeado e monitorado, para que não haja nenhuma disrupção do ponto de vista de segurança dos dados. Adicionalmente, para proteger os dados é necessário conhecimento amplo de diversas tecnologias e conceitos, pois em cada etapa existe diferentes tipos e níveis de riscos associados. A adoção de boas práticas desde o início do ciclo de vida dos dados permite que os projetos de IA sejam implementados de maneira mais rápida e segura, garantindo que os objetivos sejam atingidos sem causar danos aos negócios e/ou clientes.

Neste capítulo são descritos aspectos relacionados ao caminho dos dados, desde as suas fontes primárias até o uso em algoritmos de IA. Na seção 3.1 é detalhado o processo de extração e transformação dos dados. Conforme discutido na seção 1.4, grande parte dos vazamentos de dados ocorrem em ambientes de *analytics*. Na seção 3.2 são detalhados dois repositórios amplamente utilizados em aplicações de *analytics*: *Data Lake* e *Data Warehouse*. Os dados armazenados nestes repositórios são, então, usados para a aplicação de técnicas de IA, conforme é descrito no Capítulo 6.

3.1 Processo ETL

De acordo com o artigo Variety of data in the ETL processes in the cloud: state of the art (DIOUF; BOLY; NDIAYE, 2018), o processo ETL (Extract-Transform-Load), é responsável por integrar dados nos datawarehouses. Na fase de ETL, os dados são extraídos de várias fontes, transformados, e carregados no datawarehouse. Este é um passo obrigatório em qualquer processo de tomada de decisão baseada em dados. Os principais passos do processo são:

1. **Extração**: Coleta de dados brutos de várias fontes, priorizando a representatividade e a completude dos dados.

2. **Transformação**: Adequação dos dados por meio de limpeza (identificação e tratamento de dados ausentes, duplicados ou inconsistentes), normalização (padronização de dados em um formato comum, como unidades de medida consistentes), agregação (combinação de dados de várias fontes em uma única estrutura de dados) e engenharia de recursos (criação de novas variáveis ou características com base nos dados existentes para melhorar a análise) para facilitar análises.

3. **Carregamento**: Transferência dos dados transformados para um repositório apropriado, registrando metadados para rastreabilidade.

É fundamental considerar questões éticas e de segurança ao aplicar o ETL. Garantir a anonimização de dados sensíveis e o cumprimento das regulamentações de proteção de dados nesta etapa, pode reduzir drasticamente os riscos de vazamentos de dados.

3.2 Data Lakes e Data Warehouses

Data Lakes e *Data Warehouses* são duas soluções populares para armazenar grandes volumes de dados, conforme definições descritas a seguir.

> "A "Data Lake" is a methodology enabled by a massive data repository based on low cost technologies that improves the capture, refinement, archival, and exploration of raw data within an enterprise. A data lake contains the mess of raw unstructured or multi-structured data that for the most part has unrecognized value for the firm. - (FANG, 2015)

> "A de facto industry standard for integrating table-likedata is a Data Warehouse (DW) architecture [3], [4]. In this architecture, multiple data sources (DSs) are connected by an integration layer. This layer, commonly called extract-transform-load (ETL), runs ETL processes. They are responsible for: (1) ingesting data from DSs, (2) transforming these data into a common data model and structures, (3) cleaning and homogenizing data, (4) eliminating duplicates, and (5) uploading the final integrated data set into a central repository, which is called a Data Warehouse. - (WREMBEL, 2021)

Data Lakes e *Data Warehouses* possuem características semelhantes. Contudo, a aplicação prática dessas tecnologias demonstra claramente suas principais diferenças. Para diferenciar os termos, e facilitar o entendimento, a seguir é feita uma analogia entre esses termos.

Um *lake* (lago) é um grande repositório no qual podem existir diversas fontes de entrada. Os dados são armazenados a qualquer momento, usualmente em seu formato bruto. Nos *Data Lakes*, dados brutos provindos de diversas fontes são armazenados de forma não estruturada.

Já um *warehouse* (armazém) também pode armazenar dados obtidos de diversas fontes. Porém, ele tem uma premissa maior de organização, de forma que os dados devem passar previamente por algum tipo de identificação, organização, limpeza e agrupamento, a fim de garantir que o armazém facilite a localização de um determinado dado de maneira rápida e eficiente. Assim, *data warehouses* recebem dados do *Data Lake* e também de outras fontes, aplicam o processo ETL (extração, transformação e carregamento) sobre esses dados, e organizam e armazenam os dados usualmente segundo o modelo multidimimensional. Como resultado, os dados servem como fonte para geração de relatórios analíticos e de aprendizado de máquina.

Dentre as diferenças entre *Data Warehouses* e *Data Lakes* relacionadas no artigo (AISSI *et al.*, 2022), pode-se citar:

1. **Estrutura dos dados:** Data Warehouses geralmente armazenam dados estruturados que foram limpos e transformados para análise, enquanto Data Lakes contêm dados brutos, não estruturados ou semi-estruturados.

2. **Fonte dos dados:** Data Warehouses tendem a extrair dados de sistemas transacionais e outras fontes de dados estruturados, enquanto Data Lakes podem armazenar dados de uma variedade de fontes, incluindo fontes não estruturadas como mídias sociais, logs, sensores e outros dispositivos IoT.

3. **Processamento de dados:** Data Warehouses são projetados para oferecer suporte para consultas e relatórios pré-definidos e usam um esquema pré-definido para organizar dados. Data Lakes, por outro lado, são mais flexíveis e permitem que os dados sejam analisados de várias maneiras.

4. **Acesso aos dados:** Data Warehouses geralmente fornecem um conjunto limitado de relatórios e painéis pré-definidos, enquanto Data Lakes permitem mais flexibilidade em termos de acesso e análise de dados.

5. **Governança dos dados:** Data Warehouses tendem a ter controle mais rigoroso de governança e qualidade dos dados, enquanto Data Lakes exigem mais gerenciamento para garantir a qualidade dos dados e controles de acesso.

Do ponto de vista de *compliance*, ambos repositórios de armazenamento podem conter um número elevado de dados sensíveis. Isso demostra que a investigação da segurança de dados nestes repositórios é importante para começar a construir uma arquitetura segura para um projeto de IA.

Por exemplo, cientistas de dados podem analisar dados parcialmente anonimizados, nos quais CPFs ou nomes de clientes podem ser protegidos por meio de um processo de transformação (anonimização), que tem com objetivo de impedir a identificação do

proprietário daquele registro. Em um banco, todas as análises financeiras poderiam ser feitas usando os dados reais relativos às movimentações financeiras, porém o CPF associado a esses registros seria anônimo.

4 LEIS E REGULAMENTAÇÕES PARA PROTEÇÃO DE DADOS

Neste capítulo são detalhados aspectos relacionados às leis e às regulamentações para proteção de dados. Na seção 4.1 é feito um histórico dos marcos mais importantes. Na seção 4.2 são sumarizadas leis, regulamentos e regulamentações. Na seção 4.3 são discutidas as multas que podem ser aplicadas frente aos vazamentos de dados. Na seção 4.4 são discutidas leis específicas que se aplicam a IA.

4.1 Histórico

De acordo com (LUGATI; ALMEIDA, 2020b), as leis de proteção de dados pessoais surgiram em todo o mundo a partir da década de 1970, quando os países começaram a se preocupar com a privacidade dos dados pessoais de seus cidadãos. Desde então, houve um movimento global para criar leis de proteção de dados que abrangem coleta, armazenamento, uso e compartilhamento de dados pessoais.

O artigo também menciona os marcos mais importantes na história das leis de proteção de dados, listados a seguir:

- **1980:** A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) aprovou as Diretrizes de Proteção de Dados Pessoais, um marco importante na proteção internacional da privacidade de dados.
- **1995:** A União Europeia (UE) aprovou a Diretiva de Proteção de Dados Pessoais, que estabeleceu regras para a coleta, armazenamento e uso de informações pessoais em todos os Estados membros.
- **2000:** Os Estados Unidos aprovaram a Lei de Proteção à Privacidade Online de Crianças (COPPA), que regulamentou a coleta de informações pessoais de crianças menores de 13 anos.
- **2010:** A Comissão Europeia propôs um novo regulamento de proteção de dados para substituir a Diretiva de 1995.
- **2016:** A UE aprovou o Regulamento Geral de Proteção de Dados (GDPR), que entrou em vigor em **2018** e estabeleceu as regras para o tratamento de dados pessoais em toda a UE.
- **2018:** O Brasil aprovou a Lei Geral de Proteção de Dados (LGPD), que entrou em vigor em **2020** e regulamentou o tratamento de dados pessoais em todas as empresas que operam no Brasil.

Esses marcos legais marcaram uma evolução significativa na proteção de dados pessoais e serviram como modelo para muitos outros países em todo o mundo. A tendência global é a de que cada vez mais países adotem leis de proteção de dados pessoais para proteger a privacidade de seus cidadãos.

O gráfico ilustrado na Figura 1 mostra a evolução da busca pelo termo LGPD no Google nos últimos 5 anos. O aumento da procura ao longo dos anos comprova interesse contínuo do público brasileiro sobre esse tema desde o seu surgimento em 2018.

Figura 1 – Fonte: Extraído de Google Trends (2023)

4.2 Leis, Regulamentos e Regulamentações

GDPR significa Regulamento Geral de Proteção de Dados e é uma lei europeia que entrou em vigor em 27 de abril de 2016. Ela estabelece regras para a proteção de dados pessoais dentro da UE, incluindo a coleta, armazenamento, uso e compartilhamento dessas informações, bem como direitos dos titulares dos dados, tais como o direito de acesso, correção e exclusão de dados. A GDPR se aplica a todas as empresas que operam na UE ou lidam com dados pessoais de cidadãos europeus, independentemente de sua localização. (LUGATI; ALMEIDA, 2020a)

LGPD significa Lei Geral de Proteção de Dados e é uma lei brasileira que entrou em vigor em setembro de 2020. Ela estabelece regras para o tratamento de dados pessoais, incluindo a coleta, armazenamento, uso e compartilhamento dessas informações, bem como direitos dos titulares dos dados, tais como o direito de acesso, correção e exclusão de dados. A LGPD se aplica a todas as empresas, públicas ou privadas, que operam no Brasil e que lidam com dados pessoais. (LUGATI; ALMEIDA, 2020a)

Adicionalmente, HIPAA (Health Insurance Portability and Accountability Act), PCI (Payment Card Industry Data Security Standard) e ISO (International Organization for Standardization) são três importantes conjuntos de normas e regulamentos relacionados à proteção de dados em diferentes setores.

O HIPAA [1] é uma legislação dos Estados Unidos voltada para a área da saúde. Criada para proteger informações de saúde pessoais, ela impõe diretrizes rígidas para a privacidade e segurança de dados em ambientes médicos. Instituições de saúde, médicos e prestadores de serviços de saúde devem garantir a confidencialidade das informações do paciente, limitando o acesso apenas aos profissionais autorizados e implementando medidas de segurança robustas para proteger contra vazamentos e violações de dados.

Por sua vez, o PCI DSS [2] é um conjunto de padrões de segurança criado pelas principais empresas de cartões de crédito para proteger dados de pagamento e transações. Empresas que lidam com cartões de crédito devem estar em conformidade com o PCI DSS para garantir a proteção dos dados dos clientes durante as transações. Isso envolve a criptografia dos dados de pagamento, a manutenção de redes seguras e a implementação de práticas rigorosas de segurança para evitar roubos de informações financeiras.

Já a ISO 27001 [3] é uma norma internacional que estabelece diretrizes para o sistema de gestão da segurança da informação. Independente do setor, a ISO 27001 pode ser aplicada para ajudar as organizações a protegerem seus dados e informações sensíveis. Ela se concentra em identificar os riscos de segurança, implementar controles adequados e garantir a continuidade dos processos de negócios em caso de incidentes de segurança.

Em resumo, essas três regulamentações, HIPAA, PCI e ISO 27001, são fundamentais para a proteção de dados em seus respectivos domínios: saúde, transações financeiras e segurança da informação em geral. Assim como para a GDPR e para a LGPD, o cumprimento das normas é essencial para garantir a privacidade, integridade e disponibilidade dos dados, além de fortalecer a confiança dos clientes e parceiros nas organizações que se comprometem a proteger dados sensíveis.

4.3 Multas

O valor total das multas aplicadas pelas leis de proteção de dados em todo o mundo é difícil de ser precisamente calculado, uma vez que cada país tem suas próprias leis e regulamentações, com valores de multa diferentes. No entanto, pesquisas na internet (Google) permitiram identificar alguns dos maiores vazamentos e suas respectivas multas, os quais são detalhados na Tabela 1. Os dados exibidos nesta tabela foram coletadas em abril de 2023 a partir de diversas fontes de notícias e relatórios oficiais de órgãos reguladores de cada país, com destaque para: (i) Comissão Nacional de Proteção de Dados (CNPD) - Portugal; (ii) Information Commissioner's Office (ICO) - Reino Unido; (iii) Federal Trade Commission (FTC) - Estados Unidos; (iv) Office of the Privacy Commissioner of Canada

[1] https://www.hhs.gov/ocr/get-help-in-other-languages/portuguese.html

[2] https://www.pcisecuritystandards.org/faq/articles/Frequently$_A$sked$_Q$uestion/Does$-$PCI$-$DSS$-$apply$-$to$-$bank$-$account$-$data/

[3] https://www.27001.pt/

(OPC) - Canadá; e (v) European Data Protection Board (EDPB) - UE.

Tabela 1 – Maiores Multas por Vazamento de Dados Sensíveis

Ano	**Empresa**	**Valor da Multa (em milhões USD)**
2019	Equifax	700
2020	British Airways	230
2016	Uber	148
2018	Google (Google LLC)	57
2016	Yahoo (Altaba Inc.)	50
2019	Marriott International	23
2019	Facebook	5

A multas exibidas na Tabela 1, que juntas somam um montante de 6.6 bilhões de dólares, são um sinal claro de que as autoridades estão aplicando políticas para garantir a proteção de dados pessoais e estão dispostas a aplicar penalidades significativas a empresas que não seguem as leis e regulamentações de proteção de dados. A multas aplicadas pela LGPD podem representar 2% do faturamento da empresa ou até R$ 50 milhões por infração.

É importante destacar que o objetivo das multas não é apenas punir as empresas, mas também incentivar a adoção de práticas mais seguras e responsáveis em relação aos dados pessoais.

4.4 Leis específicas para IA

Na seção 4.4.1 é detalhada a primeira regulamentação específica para IA. Na seção 4.4.2 são descritos os quatro níveis de risco definidos pelo marco regulatório do parlamento europeu.

4.4.1 Primeira regulamentação

A primeira regulamentação específica para IA é definida pela comissão europeia. Desde 2021, essa comissão tem trabalhado na construção de um modelo para definir regras para o uso de dados de cidadãos europeus em projetos de IA, conforme pode ser observado na citação a seguir.

> "AI technologies are expected to bring a wide array of economic and societal benefits to a wide range of sectors, including environment and health, the public sector, finance, mobility, home affairs and agriculture. They are particularly useful for improving prediction, for optimising operations and resource allocation, and for personalising services.1 However, the implications of AI systems for fundamental rights protected under the EU Charter of Fundamental Rights, as well as the safety risks for users when AI technologies

> are embedded in products and services, are raising concern. Most notably, AI systems may jeopardise fundamental rights such as the right to non-discrimination, freedom of expression, human dignity, personal data protection and privacy."

Essa citação é um trecho do documento Artificial intelligence act (PARLIAMENT, 2022), criado pelo parlamento europeu, no qual são discutidos pilares fundamentais para controle da AI. Pretende-se finalizar as negociações com os países do bloco europeu até o final de 2023.

A prioridade do Parlamento é garantir que os sistemas de IA utilizados na UE sejam seguros, transparentes, rastreáveis, não discriminatórios e respeitadores do ambiente. Ainda de acordo com as prioridades do Parlamento, os sistemas de IA devem ser supervisionados por pessoas, em vez de serem automatizados, para evitar resultados prejudiciais. O Parlamento quer também estabelecer uma definição uniforme e neutra em termos tecnológicos para a IA, de modo a ser aplicada em futuros sistemas de IA.

O regulamento de IA proposto garante que os europeus possam confiar no que a IA tem a oferecer. Embora a maioria dos sistemas de IA represente um risco limitado e possa contribuir para a solução de muitos desafios sociais, certos sistemas de IA criam riscos que devem ser analisados para evitar resultados indesejáveis. Por exemplo, muitas vezes não é possível descobrir por que um sistema de IA tomou uma decisão ou previsão e realizou uma determinada ação. Portanto, pode ser difícil avaliar se alguém foi injustamente prejudicado, como em uma decisão de contratação ou em uma inscrição para um esquema de benefício público.

As regras propostas vislumbram:

- Lidar com riscos criados especificamente por aplicativos de IA.
- Propor uma lista de aplicativos de alto risco.
- Definir requisitos claros para sistemas de IA para aplicações de alto risco.
- Definir obrigações específicas para usuários de IA e provedores de aplicativos de alto risco.
- Propor uma avaliação de conformidade antes do sistema de IA entrar em serviço ou ser colocado no mercado.
- Propor a aplicação após tal sistema de IA ser colocado no mercado.
- Propor uma estrutura de governação em nível europeu e nacional.

4.4.2 Níveis de risco do marco regulatório

O marco regulatório do parlamento europeu define quatro níveis de risco em IA, conforme ilustrado na Figura 2. Esses riscos são: (i) risco inaceitável; (ii) alto risco; (iii) risco limitado; e (iv) risco mínimo ou nenhum. Cada um desses riscos é descrito a seguir.

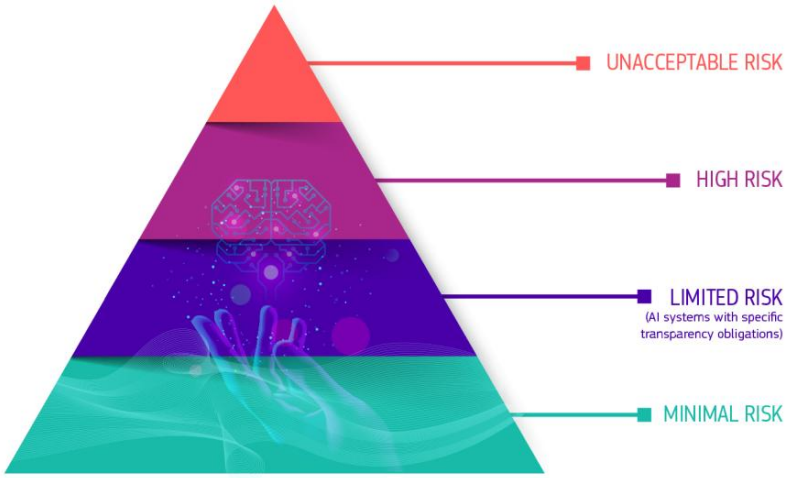

Figura 2 – Pirâmide de Riscos - Fonte: (PARLIAMENT, 2022)

Os sistemas de IA de **risco inaceitável** são sistemas considerados uma ameaça para as pessoas e serão proibidos. Estes sistemas incluem:

- Manipulação cognitivo-comportamental de pessoas ou grupos vulneráveis específicos, como brinquedos ativados por voz que incentivam comportamentos perigosos nas crianças.
- Pontuação social: classificação de pessoas com base no comportamento, estatuto socioeconômico e caraterísticas pessoais.
- Sistemas de identificação biométrica em tempo real e à distância, como o reconhecimento facial.

Podem ser permitidas algumas exceções. Por exemplo, os sistemas de pós-identificação biométrica à distância, nos quais a identificação ocorre após um atraso significativo, são permitidos somente para a repressão de crimes graves e após a aprovação do tribunal.

Os **riscos elevados**, por sua vez, referem-se aos sistemas de IA que afetam negativamente a segurança ou os direitos fundamentais. Eles são divididos em duas categorias. A primeira categoria refere-se aos sistemas de IA que são utilizados em produtos

abrangidos pela legislação da UE em matéria de segurança dos produtos. Isto inclui brinquedos, aviação, automóveis, dispositivos médicos e elevadores.

A segunda categoria diz respeito aos sistemas de IA que se enquadram em uma ou mais das seguintes oito áreas específicas: (i) identificação biométrica e categorização de pessoas singulares; (ii) gestão e funcionamento de infraestruturas essenciais; (iii) educação e formação profissional; (iv) emprego, gestão dos trabalhadores e acesso ao trabalho por conta própria; (v) acesso e usufruto de serviços privados essenciais e de serviços e benefícios públicos; (vi) aplicação da lei; (vii) gestão da migração, do asilo e do controlo das fronteiras; e (viii) assistência na interpretação jurídica e na aplicação da lei. Esses sistemas têm que ser obrigatoriamente registados em uma base de dados da UE.

Todos os sistemas de IA de risco elevado serão avaliados tanto antes de serem colocados no mercado como durante todo o seu ciclo de vida.

Quanto ao **risco limitado**, ele se refere aos sistemas de IA com obrigações específicas de transparência. Ao usar sistemas de IA, como *chatbots*, os usuários devem estar cientes de que estão interagindo com uma máquina para que possam tomar uma decisão informada de continuar ou recuar.

Finalmente, o marco regulatório do parlamento europeu permite o uso gratuito de IA de **risco mínimo ou nenhum**. Isso inclui aplicativos como videogames habilitados para IA ou filtros de spam. A grande maioria dos sistemas de IA atualmente usados na UE se enquadra nessa categoria.

Para complementar as discussões realizadas nesta seção, é necessário contextualizar a IA generativa, tal como o ChatGPT[4]. Ela tem que cumprir os requisitos de transparência descritos a seguir. O conteúdo do que foi gerado pela IA deve ser divulgado. Adicionalmente, o modelo utilizado para gerar o conteúdo deve ser concebido de forma a evitar a geração de conteúdos ilegais. Outro requisito consiste em publicar apenas os resumos de dados protegidos por direitos de autor utilizados para a formação.

[4] https://chat.openai.com/

5 COMO IDENTIFICAR E PROTEGER DADOS SENSÍVEIS

Neste capítulo são descritos os resultados relacionados ao aspecto teórico investigado no trabalho, o qual é relacionado a como identificar e proteger dados sensíveis. Na seção 5.1 são categorizados os dados que podem ser enquadrados como informações de identificação pessoal. Nas seções 5.2 e 5.3 são descritas técnicas e ferramentas para identificação e proteção de dados sensíveis. Na seção 5.4 são discutidas estratégias para a proteção que podem ser adotadas pelos cientistas de dados em projetos de IA.

5.1 Informações de identificação pessoal

PII (Informações de Identificação Pessoal) são dados que podem ser usados para identificar, contatar ou localizar um indivíduo específico, seja sozinho ou em combinação com outras informações. As categorias de PII podem variar dependendo do contexto e das regulamentações específicas, mas geralmente incluem as seguintes:

- **Informações de identificação básicas:** Incluem nome completo, data de nascimento, gênero, número de identificação, número de passaporte e outros detalhes que são usados para identificar diretamente uma pessoa.
- **Informações de contato:** Abrange informações de contato pessoais, como endereço residencial, número de telefone, endereço de e-mail e outras formas de comunicação direta com o indivíduo.
- **Informações financeiras:** Inclui dados relacionados às contas bancárias, números de cartão de crédito, histórico de transações financeiras e outras informações financeiras que podem ser usadas para fins de pagamento e identificação em transações.
- **Informações de saúde:** Estão relacionadas à saúde física e mental do indivíduo e podem incluir histórico médico, condições médicas, prescrições, exames de laboratório e outras informações de saúde.
- **Informações de emprego:** Referem-se aos detalhes sobre o emprego de uma pessoa, como histórico de empregos, salário, número de funcionário, benefícios, informações de recuross humanos e outros dados relacionados ao emprego.
- **Informações educacionais:** Engloba dados sobre a educação de uma pessoa, como histórico escolar, diplomas, cursos realizados e outras informações educacionais.
- **Informações de identificação biométrica:** Inclui dados biológicos ou comportamentais usados para identificar uma pessoa de forma única, como impressões digitais, reconhecimento facial, íris ou voz.

- **Informações de localização:** Referem-se à localização geográfica de uma pessoa, seja por meio de coordenadas GPS, endereço físico ou outra forma de rastreamento.
- **Outras informações identificáveis:** Outros dados que, quando combinados com outras informações, podem levar à identificação de uma pessoa específica, como dados de redes sociais, preferências pessoais e interesses, dentre outros.

Adicionalmente, é possível agrupar os PIIs em 2 grandes grupos: dados sensíveis e dados pessoais. Um **dado sensível** é qualquer informação que, se divulgada, pode resultar em discriminação, preconceito, dano emocional ou financeiro a um indivíduo. Dados sensíveis são considerados especialmente protegidos devido ao seu potencial para causar impacto significativo na privacidade e na vida das pessoas.

Exemplos de dados sensíveis incluem informações médicas, origem racial ou étnica, convicções religiosas ou filosóficas, orientação sexual, informações genéticas e opiniões políticas, dentre outros. A proteção adequada desses dados é essencial para garantir a privacidade e evitar possíveis consequências negativas.

Um **dado pessoal**, por sua vez, é qualquer informação relacionada a uma pessoa física identificada ou identificável. Isso inclui qualquer informação que, sozinha ou em combinação com outras informações, possa identificar uma pessoa específica. Exemplos de dados pessoais incluem nome, endereço, número de identificação, informações de contato, dados biométricos e informações financeiras.

Os dados pessoais são amplamente coletados e usados em várias atividades cotidianas, como compras online, acesso a serviços, comunicação e interações com empresas e organizações. Devido à sua natureza identificável, os dados pessoais também exigem proteção adequada para garantir a privacidade, a segurança e o cumprimento das leis de proteção de dados.

5.2 Técnicas para proteção de dados sensíveis

Atualmente existem várias técnicas e práticas recomendadas para proteger dados sensíveis em projetos de IA. Dentre as mais utilizados, pode-se citar:

- **Anonimização e pseudonimização:** Remove ou substitui informações pessoais identificáveis, como nomes, endereços e números de identificação, por identificadores não pessoais para garantir que os dados não possam ser vinculados diretamente a uma pessoa específica.
- **Criptografia:** Aplica técnicas de criptografia para proteger os dados sensíveis durante o armazenamento, transmissão e processamento, garantindo que apenas pessoas autorizadas possam acessar os dados.

- **Minimização de dados:** Coleta apenas os dados estritamente necessários para o objetivo do projeto de IA, reduzindo a quantidade de informações sensíveis que precisam ser protegidas.
- **Acesso restrito:** Limita o acesso aos dados sensíveis apenas a pessoas autorizadas por meio de controle de acesso, autenticação e autorização adequados.
- **Segurança de rede:** Protege as redes utilizadas no projeto de IA para evitar acesso não autorizado e interceptação de dados, utilizando firewalls, VPNs e outras medidas de segurança para fortalecer a infraestrutura de rede.
- **Treinamento em ambiente seguro:** Envolve a utilização de técnicas de treinamento federado, as quais permitem o treinamento de modelos sem que os dados reais sejam compartilhados entre os servidores, com o objetivo de garantir que os dados sensíveis não sejam expostos acidentalmente.
- **Avaliação de riscos e impacto de privacidade:** Consiste na realização de avaliações de risco e impacto de privacidade para identificar e mitigar potenciais ameaças à segurança dos dados sensíveis e às informações pessoais dos usuários.
- **Auditoria e monitoramento:** Implementa sistemas de auditoria e monitoramento para rastrear o acesso e uso dos dados sensíveis, permitindo detectar comportamentos suspeitos ou atividades não autorizadas.
- **Conformidade com regulamentações:** Especifica a garantia de conformidade com as leis e regulamentos de privacidade de dados relevantes ao projeto de IA.
- **Educação e conscientização:** Propõe o treinamento de todos os membros da equipe do projeto de IA sobre as melhores práticas de proteção de dados e a importância da privacidade.

Essas técnicas devem ser adaptadas ao contexto específico de cada projeto de IA e devem ser constantemente revisadas e atualizadas à medida que novas ameaças e desafios surgirem. A segurança dos dados sensíveis é uma responsabilidade contínua e compartilhada por todos os envolvidos no projeto.

5.3 Ferramentas para identificação de dados sensíveis

Além das técnicas mencionadas na seção anterior, é muito importante contar com o apoio de ferramentas que facilitem o processo de governança dos dados. Existem 2 grandes categorias de ferramentas para identificação de dados sensíveis. A primeira atua diretamente nas fontes primárias de dados (SGBDs e DataLakes), e a segunda oferecem uma abordagem com foco na anonimização integrada aos *data pipelines* em ambientes de *cloud*.

Na primeira categoria existe um grande desafio devido à falta de documentação do modelo da base de dados, que na maioria das vezes, impossibilita identificar onde estão as localizadas as informações sensíveis em um banco de dados.

O Delphix [1], uma ferramenta criada há 15 anos por uma empresa localizada no Vale do Silício, possibilita uma varredura em bases de dados (estruturadas e não-estruturadas), e também em arquivos como JSON, XML e CSV, utilizando conjuntos de expressões regulares para identificar a catalogar colunas com dados sensíveis dentro do contexto da LGPD (além de outras regulamentações). A solução também oferece recursos para fazer o mascaramento de dados usando dados fictícios, porém realísticos, mantendo a integridade referencial em múltiplas fontes de dados.

Como alternativa de baixo custo, o ARX [2] (Anonymization & Privacy-preserving Research eXtension) é um software de código aberto projetado para a anonimização e proteção da privacidade em conjuntos de dados, particularmente em contextos de pesquisa onde a confidencialidade dos dados sensíveis deve ser preservada. Este software é usado para aplicar técnicas de anonimização em dados pessoais, de forma a torná-los seguros para uso em análises, sem comprometer a privacidade das pessoas envolvidas.

Outras abordagens mais modernas já contam com o apoio da própria IA para identificar e proteger dados sensíveis. O projeto Presidio [3] da Microsoft é um deles, e faz uma mistura de técnicas tradicionais com o uso de algorítimos de *machine learning* para apoiar projetos de *compliance*.

5.4 Estratégias para Cientistas de Dados

Existem várias estratégias que um cientista de dados pode seguir diante de requisitos de privacidade de dados extremamente desafiadores. Com base nas pesquisas realizadas durante o desenvolvimento deste trabalho, são propostas 7 estratégias que podem ajudar na proteção dos dados em projetos de IA. Essas estratégias são descritas a seguir.

1. **Excluir o conjunto de dados após o treinamento do modelo**

   Durante o projeto, dados anotados podem ser usados. Apenas uma cópia do conjunto de dados pode existir. No entanto, depois que o modelo de aprendizado de máquina for treinado, o cientista de dados deve excluir o conjunto de dados completo.

   Excluir o conjunto de dados significa que, se o projeto for retomado no futuro, o cientista de dados precisará anotar novamente os dados. Por outro lado, se todos os dados forem realmente excluídos, não há como eles vazarem.

[1] https://delphix.com
[2] arx.deidentifier.org
[3] https://microsoft.github.io/presidio/

2. **Mascarar dados sensíveis**

   Uma outra técnica é o mascaramento de dados. Campos como e-mail, nome, endereço, números de documentos (CPF, CNPJ, PIS, etc), e outras informações consideradas pessoais e/ou sensíveis, podem ser anonimizadas por meio desta técnica, que consiste em trocar dados reais por fictícios.

   Ao aplicar algoritmos para proteger os dados usando mascaramento (anonimização), é de suma importância que durante a transformação dos dados seja mantida a integridade referencial. A integridade referencial é um termo que se refere à lógica de relacionamento entre os campos de diferentes tabelas de um banco de dados. Por exemplo, caso o campo de CPF seja modificado na tabela A, é necessário aplicar a mesma modificação em todos os campos correlacionados (tabelas B, C, D...), garantindo assim que a semântica de negócio seja mantida.

   Caso seja bem aplicada, essa técnica permite entregar dados de qualidade e seguros para modelos de aprendizado de máquina. Caso ocorra algum tipo de vazamento desses dados, não será possível fazer a re-identificação.

3. **Armazenar apenas identificadores que podem ser usados para reconstruir dados**

   Ao invés de gerar endereços de e-mail fictícios, é possível converter os dados para *hash* e gerar um rótulo que permita a reconstrução dos dados. Em comparação ao mascaramento de dados, o uso de identificadores (IDs) torna o processo mais rápido. Contudo, essa técnica não é 100% segura pois a partir de outras bases de dados que contenham endereços de e-mail, um hacker pode fazer o *hash* e comparar com os endereços do conjunto de dados.

4. **Utilizar criptografia homomórfica**

   A criptografia homomórfica é uma técnica criptográfica avançada que permite a realização de operações matemáticas em dados criptografados sem a necessidade de descriptografia prévia. Essa capacidade de processar dados mantidos em um estado criptografado preserva a privacidade e a confidencialidade dos dados, tornando-a uma ferramenta relevante em diversos contextos.

   Existem três propriedades principais da criptografia homomórfica:

   - **Homomorfismo Aditivo:** Possibilita a execução de operações de adição em dados criptografados. Por exemplo, a soma de dois números criptografados resulta em um valor criptografado que, quando descriptografado, corresponde à soma dos números originais.
   - **Homomorfismo Multiplicativo:** Permite a realização de operações de multiplicação em dados criptografados. Analogamente à propriedade aditiva, a

multiplicação de dois números criptografados resulta em um valor criptografado cuja descriptografia corresponde ao produto dos números originais.

- **Homomorfismo Completo**: Esta é uma propriedade avançada que habilita a execução de qualquer operação matemática nos dados criptografados, não se limitando apenas à adição e multiplicação.

5. **Remover dados confidenciais que não são importantes para o modelo**

   Dados que não importantes devem ser removidos. Por exemplo, a data de nascimento pode não fazer nenhuma diferença no resultado do treinamento de um modelo. Portanto, ela pode ser removida.

   Em situações nas quais a diferença de desempenho do modelo seja pouco afetada pela ausência de um determinado campo sensível, é sempre mais prudente evitar a exposição do mesmo por meio de sua remoção.

6. **Reduzir os dados confidenciais**

   A técnica de redução de dados é um processo usado na análise de dados e na ciência de dados para diminuir a quantidade de variáveis a serem analisadas em um conjunto de dados, enquanto se mantém uma representação significativa e útil dos dados originais. Essa técnica é frequentemente aplicada para tornar os dados mais gerenciáveis, reduzir a complexidade computacional de análises subsequentes e eliminar informações irrelevantes ou redundantes. Adicionalmente, ela também pode ser uma forte aliada para a proteção de dados sensíveis.

   Existem duas abordagens principais para a redução de dados. A primeira, conhecida como seleção de características (*Feature Selection*), envolve a escolha de um subconjunto das características ou variáveis originais de um conjunto de dados, mantendo apenas aquelas que são mais relevantes para a análise ou tarefa em questão. (LI *et al.*, 2017) A segunda, conhecida como redução de dimensionalidade, visa reduzir o número de variáveis (ou dimensões) nos dados, transformando-os em um espaço de menor dimensão, geralmente por meio de técnicas matemáticas. (HARRISON, 2019)

   A escolha entre seleção de características e redução de dimensionalidade depende do problema específico em questão. Em alguns casos, ambas as técnicas podem ser usadas em conjunto para otimizar a representação dos dados. No entanto, é importante lembrar que a redução de dados deve ser realizada com cuidado, pois a perda de informações pode afetar a qualidade das análises e dos modelos resultantes. Portanto, é essencial considerar os objetivos da análise e o impacto potencial da redução de dados no contexto do problema.

7. **Manter os dados confidenciais em um silo**

   Também é possível manter os dados sensíveis em um repositório seguro onde os pesquisadores não podem acessá-los diretamente, mas podem submeter experimentos e realizar testes estatísticos.

   Por exemplo, o Serviço Nacional de Saúde (NHS) da Inglaterra criou um programa piloto chamado OpenSAFELY[4], que permite aos pesquisadores usar registros de saúde de 58 milhões de pessoas sem nunca vê-los. Os usuários podem escrever código e enviá-lo na plataforma (que pode ser baixado como um repositório do Github), sem precisar visualizar os registros brutos. Todas as interações com os dados são registradas e os projetos aprovados são listados no sítio da OpenSAFELY.

   Para prover as funcionalidades mencionadas anteriormente, o OpenSafely usa um conjunto de tabelas em camadas e não provê aos pesquisadores acesso para executar consultas simples ao banco de dados e ver os dados brutos. A arquitetura é centrada em uma plataforma de análise segura na qual o código é executado.

[4] https://docs.opensafely.org/

6 IA PARA IDENTIFICAÇÃO DE CAMPOS SENSÍVEIS

Os resultados teóricos descritos no Capítulo 5 mostram que é possível preservar a privacidade dos dados através de diversas técnicas. Neste capítulo, é proposta uma abordagem prática diferenciada, onde técnicas de IA são avaliadas com objetivo de verificar a sua efetividade na identificação de dados sensíveis. Em detalhes, foi treinado um algoritmo para verificar a sua capacidade na identificação de PIIs.

Foram utilizadas técnicas que misturam o uso de expressões regulares (REGEX) na etapa de rotulagem dos dados, scikit-learn para efetuar a classificação usando um vetorizador TF-IDF e um modelo de *Logistic Regression* (LR). Na seção 6.1 são descritos como os dados foram preparados para o treinamento do modelo. Em seguida, a seção 6.2, explica o processo de classificação e rotulagem dos dados através do uso de expressões regulares. Na seção 6.3, um detalhamento dos motivos que levaram à escolha do LR e um detalhamento das principais etapas do ciclo de treinamento, validação e testes. Na seção 6.4, uma explicação do código utilizado para treinar e validar o modelo proposto, e por fim, na seção 6.5, um resumo dos resultados obtidos.

6.1 Dados para treinamento do modelo

Com objetivo de criar uma metodologia flexível, e que possa ser replicada em diferentes formatos de tabelas, optou-se pela realização da análise em nível de dado. Assim, por meio de um conjunto de expressões regulares, os dados são analisados e classificados de acordo com o tipo de dado identificado. Essa escolha considera o fato de que nem sempre os metadados, ou seja, os nomes das colunas, permitem identificar a presença de dados sensíveis. Por exemplo, diversos sistemas usam códigos, *strings* complexas e muitas vezes sem sentido, o que faz com que o processo de catalogação dos dados seja muito complexo e garanta baixa assertividade.

Para essa tarefa, levou-se em consideração campos que fazem parte do escopo da LGPD, com destaque para: (i) CNPJ; (ii) CPF; (iii) Nome Completo; e (iv) Estado. O primeiro passo consistiu na criação de uma tabela de controle para armazenar os tipos de dados a serem identificados. Em Python, foi criado um *dataset* de forma dinâmica, com duas colunas, uma contendo o “tipo de dado” e outra com a expressão regular a ser aplicada (Figura 3). Essa técnica tem como vantagem o fato de permitir acrescentar facilmente novos tipos de dados a serem catalogados.

```
# Função para criar a tabela "tipos_dados_sensiveis"
def criar_tipos_dados_sensiveis():
    tipos_dados_sensiveis_data = {
        'regex': [
            r'(?i)[A-Za-z]+ [A-Za-z]+',
            r'(?i)\d{3}\.\d{3}\.\d{3}-\d{2}',
            r'^(?!(\d)\1{13})\d{14}$',
            r'(?i)\(\d{2}\) \d{4,5}-\d{4}',
            r'(?i)Rua .+|Avenida .+|Travessa .+',
            r'\d{5}-\d{3}',
            r'^(?:AC|AL|AP|AM|BA|CE|DF|ES|GO|MA|MT|MS|MG|PA|PB|PR|PE|PI|RJ|RN|RS|RO|RR|SC|SP|SE|TO)$'
            # Adicione mais expressões regulares para outros tipos de dados sensíveis
        ],
        'tipo_dado': ['nome', 'cpf', 'cnpj', 'telefone', 'logradouro', 'cep', 'estado']
        # Adicione mais tipos de dados sensíveis aqui
```

Figura 3 – Expressões Regulares LGPD

Como não existem *datasets* públicos com dados pessoais, foi necessário utilizar a técnica de aumento de dados. Para tanto, foi utilizado um *dataset* obtido no portal de dados abertos do Governo Federal[1]. Esse *dataset* contém dados com nomes de sócios de empresas. Destaca-se que dados relacionados aos números de CNPJ e CPF, dentre outros, são considerados dados sensíveis e, portanto, encontram-se parcialmente mascarados.

O *dataset* possui 200 mil registros e, para fazer o aumento de dados, foi utilizada a linguagem Python. Colunas com CPF, CNPJ e Estado foram preenchidas com dados fictícios, porém válidos. O código usado para essa finalidade é ilustrado na Figura 4. Na Figura 5 é ilustrada uma amostra dos dados disponibilizados no *dataset* original. Na Figura 6 são ilustrados os mesmos nomes, porém acompanhados de números de documentos gerados aleatoriamente.

```
        return cpf_str

    # Create a new column with fictitious CPFs
    new_cpf_column = [gerar_cpf_hash() for _ in range(len(df))]
    df.insert(0, "SOCIO_CPF", new_cpf_column)

    # Create a new column with fictitious CNPJ numbers
    new_cnpj_column = [generate_fake_cnpj() for _ in range(len(df))]
    df.insert(0, "CNPJ", new_cnpj_column)
```

Figura 4 – Criação de Dados Fictícios

[1] https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica—cnpj

	CNPJ_BASICO	COD_IDENT_SOCIO	NOME_SOCIO	CNPJCPF_SOCIO	QUALIFIC_SOCIO	DATA_INGRESSO	PAIS	REP_LEGAL	NOME_REP_LEGAL	QUALIFIC_REP_LEGAL	COD
0	1879005	2	LUCIANO FONSECA	***477633**	16	20050912	NaN	***000000**	NaN	0	
1	1879008	2	ERNESTO ODONE ALVES CUNHA	***594908**	49	19970513	NaN	***000000**	NaN	0	
2	1879008	2	ELAINE APARECIDA DE ALMEIDA CUNHA	***989406**	22	20060116	NaN	***000000**	NaN	0	

Figura 5 – DataSet Original

	CNPJ	SOCIO_CPF	NOME_SOCIO	ESTADO
0	19081369000197	498.148.815-70	LUCIANO FONSECA	RN
1	53409234000105	260.119.945-80	ERNESTO ODONE ALVES CUNHA	RJ
2	52722574000124	118.626.350-48	ELAINE APARECIDA DE ALMEIDA CUNHA	PE
3	11773746000189	336.031.584-74	EMERSON SANTIAGO	AC

Figura 6 – DataSet Dados Fictícios

6.2 Classificação dos Dados utilizando REGEX

Na primeira etapa do processo de identificação dos dados, o algoritmo faz uma varredura na tabela de origem em busca de dados sensíveis. Ele promove uma rotulagem dinâmica, seguindo a seguinte lógica:

1 - Todas as expressões regulares da tabela de controle são aplicadas em cada uma das colunas com os dados de origem. Para cada uma das expressões regulares, é gerada uma nova coluna para armazenar o resultado da análise. Essa coluna é preenchida com o tipo de dado identificado. Caso o dado não pertença àquela categoria, o campo é preenchido com zero.

2 - O formato de nome das colunas criadas contém o prefixo “chk_” seguido pelo nome da coluna da tabela de origem seguido pelo nome da expressão regular. O formato permite de forma lógica e visual saber qual processo de análise foi aplicado.

Caso existam n expressões regulares e o *dataset* a ser analisado possua p colunas, o resultado da análise é um *dataset* com $n*p$ colunas. Por exemplo, para 5 expressões e 10 colunas, é gerado um *dataset* com 50 colunas.

	CNPJ	SOCIO_CPF	NOME_SOCIO	ESTADO	chk_CNPJ_nome	chk_CNPJ_cpf	chk_CNPJ_cnpj	chk_CNPJ_telefone	chk_
0	89094553000180	345.825.830-24	LUCIANO FONSECA	AL	0	0	cnpj	0	
1	300438000130	591.784.739-85	ERNESTO ODONE ALVES CUNHA	CE	0	0	0	0	
2	56321179000190	767.234.980-54	ELAINE APARECIDA DE ALMEIDA CUNHA	PR	0	0	cnpj	0	
3	71685006000196	323.912.364-92	EMERSON SANTIAGO	AC	0	0	cnpj	0	
4	26029767000192	851.892.490-18	DANIELA COSTA SANTIAGO	PR	0	0	cnpj	0	

Figura 7 – Dados sensíveis rotulados

Na segunda etapa, o algoritmo remove todas as colunas novas que foram criadas com o prefixo “chk_” e que contém somente 0 (zeros). Para essas colunas, verificou-se que não havia nenhuma correspondência positiva para a existência de dados sensíveis.

Como resultado, é gerado um *dataset* contendo somente as correspondências positivas para o teste realizado com as expressões regulares conforme Figura 8. Neste ponto, é possível saber exatamente o tipo de dado sensível foi encontrado em cada coluna.

	CNPJ	SOCIO_CPF	NOME_SOCIO	ESTADO	chk_CNPJ_cnpj	chk_SOCIO_CPF_cpf	chk_NOME_SOCIO_nome
0	89094553000180	345.825.830-24	LUCIANO FONSECA	AL	cnpj	cpf	nome
1	300438000130	591.784.739-85	ERNESTO ODONE ALVES CUNHA	CE	0	cpf	nome
2	56321179000190	767.234.980-54	ELAINE APARECIDA DE ALMEIDA CUNHA	PR	cnpj	cpf	nome
3	71685006000196	323.912.364-92	EMERSON SANTIAGO	AC	cnpj	cpf	nome
4	26029767000192	851.892.490-18	DANIELA COSTA SANTIAGO	PR	cnpj	cpf	nome

Figura 8 – Dataset com rótulos filtrado

Na sequência, o algoritmo converte os nomes das expressões regulares em números usando um dicionário de dados, conforme Figura 9.

	CNPJ	SOCIO_CPF	NOME_SOCIO	ESTADO	chk_CNPJ_cnpj	chk_SOCIO_CPF_cpf	chk_NOME_SOCIO_nome	chk_ESTADO_estadc
0	22543185000130	830.073.309-47	LUCIANO FONSECA	RR	1	2	3	4
1	46678135000127	408.159.539-92	ERNESTO ODONE ALVES CUNHA	MS	1	2	3	4
2	33500380000186	849.654.497-46	ELAINE APARECIDA DE ALMEIDA CUNHA	PR	1	2	3	4
3	58964348000177	512.157.342-12	EMERSON SANTIAGO	RN	1	2	3	4
4	80944375000134	357.650.736-11	DANIELA COSTA SANTIAGO	MA	1	2	3	4

Figura 9 – Dataset com rótulos convertidos usando dicionário

```
import pandas as pd
pd = pd.DataFrame(df_filtered)

# Converting the codes to appropriate categories using a dictionary
my_dict = {
    'cnpj':'0',
    'cpf':'1',
    'nome':'2',
    'estado':'3'
}

# Iterate over columns and replace values
for col in pd.columns:
    if col.startswith('chk_'):
        pd[col] = pd[col].apply(lambda x: my_dict.get(x, x))

pd.head()
```

Figura 10 – Dicionário Dados

Na última etapa, o algoritmo combina os dados em um novo *dataframe* contendo 3 colunas: (i) uma para armazenar o texto original; (ii) outra com o rótulo em formato de *string* do tipo de dado encontrado; (iii) e uma última contendo uma *tag* em formato numérico, a ser utilizada no treinamento do modelo. Esta *tag* consiste no resultado da conversão de uma *string* de identificação para o tipo de dado encontrado. No caso, cpf, cnpj, nome e estado são transformados para os códigos de classificação 0, 1, 2 e 3, respectivamente.

	Data	Classifier	Tag
0	89094553000180	cnpj	0
1	300438000130	0	0
2	56321179000190	cnpj	0
3	71685006000196	cnpj	0
4	26029767000192	cnpj	0
...	...	...	...
295	55034284000180	cnpj	0
296	49864886000150	cnpj	0
297	46162066000191	cnpj	0
298	73015988000120	cnpj	0
299	99408260000144	cnpj	0

Figura 12 – Dados, Classificadores e Tag

```python
# Create a new DataFrame for the combined dataset
combined_data = []
combined_classifier = []

# Iterate through each pair of data and classifier columns
for data_col, classifier_col in zip(data_dataset.columns, classifier_dataset.columns):
    data_values = data_dataset[data_col]
    classifier_values = classifier_dataset[classifier_col]

    # Add the values to the combined lists
    combined_data.extend(data_values)
    combined_classifier.extend(classifier_values)

# Create a new DataFrame from the combined lists
combined_dataset = pd.DataFrame({'Data': combined_data, 'Classifier': combined_classifier})

# Print or use the combined dataset
print(combined_dataset)
```

Figura 11 – Código para geração do dataset combinado

Esses são os principais passos realizados no código para preparar o conjunto de dados que foi utilizado como fonte para o treinamento e validação do modelo proposto.

6.3 A escolha do modelo de IA

Os principais argumentos para escolha da metodologia são descritos nos 6 tópicos a seguir, onde o foco principal foi validar a efetividade da regressão logística (CARROLL; PEDERSON, 1993) para identificação dos dados sensíveis.

1. **Precisão na Rotulagem:** A utilização de expressões regulares (REGEX) permite identificar padrões específicos nos dados de maneira precisa e eficaz. Isso é especialmente útil quando os dados possuem estruturas previsíveis, como datas, endereços de email, números de telefone, etc. A combinação de REGEX com regras de rotulagem customizadas oferece maior controle sobre o processo de rotulagem, resultando em rótulos de alta qualidade.

2. **Eficiência e Velocidade:** As expressões regulares são altamente otimizadas para busca de padrões em grandes volumes de texto, tornando o processo de rotulagem mais eficiente em termos de tempo de execução. Isso é crucial, especialmente quando se lida com grandes conjuntos de dados.

3. **Adaptabilidade a Diferentes Tarefas:** A flexibilidade das REGEX permite que a metodologia seja adaptada a diversas tarefas de rotulagem, independentemente do domínio. Isso significa que a mesma estrutura de REGEX pode ser aplicada a diferentes conjuntos de dados, simplificando o processo de rotulagem em uma variedade de contextos.

4. **Aproveitamento de Características Textuais:** O uso de um vetorizador TF-IDF permite que informações textuais relevantes sejam extraídas e representadas em um formato numérico. Isso possibilita a utilização de técnicas de aprendizado de máquina, como Logistic Regression, que são altamente eficazes na classificação de texto, em especial para campos contendo informações como nomes de pessoas.

5. **Interpretabilidade do Modelo:** O modelo Logistic Regression é conhecido por sua interpretabilidade. Isso significa que é possível compreender como as palavras e características influenciam a classificação final, o que é fundamental em cenários nos quais a explicação do processo de decisão é importante.

6. **Baixa Complexidade de Implementação:** A integração do scikit-learn com REGEX e Logistic Regression (CARROLL; PEDERSON, 1993) é relativamente simples e requer menos recursos computacionais em comparação com modelos de PLN mais complexos, tornando-a uma opção atraente para projetos com recursos limitados.

A sequência de passos desenvolvida neste trabalho para a construção do modelo de IA baseado em Regressão Logística para reconhecimento de dados no escopo da LGPD é descrita a seguir.

1. O arquivo SOCIOS_02.mixed.csv é carregado usando Pandas.
2. Os dados são identificados e categorizados (rotulados) usando expressões regulares.
3. O *dataset* é dividido em conjunto de treinamento, teste e validação.
4. Na etapa de pré-processamento dos dados, os textos (dados sensíveis) são classificados usando regressão logística.
5. Os vetores de entrada são criados após a codificação dos *tokens*.
6. Os rótulos (*labels*) para a tarefa binária (tipo de dado sensível) são definidos.
7. O classificador binário é treinado utilizando Regressão Logística.
8. O resultado é avaliado usando o conjunto de testes.
9. As métricas de desempenho, como precisão, recall e F1-score são validadas.

6.4 Descrição do Código

O código a seguir realiza várias etapas de processamento de dados e treinamento de modelo usando Python e bibliotecas do scikit-learn. O objetivo é carregar um conjunto de dados, criar um modelo de Regressão Logística (CARROLL; PEDERSON, 1993) usando recursos TF-IDF e avaliar o desempenho do modelo. Além disso, o modelo treinado e o vetorizador TF-IDF são salvos em arquivos para uso futuro.

```
import pandas as pd
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
import joblib

# Carrega o conjunto de dados
data = pd.read_csv('/content/sample_data/SOCIOS_02.mixed.csv', sep=";")

# Divide o conjunto de dados em treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(data['Data'], data['Tag']
```

```
# Extrai recursos TF-IDF dos dados de texto
tfidf_vectorizer = TfidfVectorizer(max_features=1000)
X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)

# Treina um modelo de Regressao Logistica
clf = LogisticRegression()
clf.fit(X_train_tfidf, y_train)

# Avalia o modelo
y_pred = clf.predict(X_test_tfidf)
print(classification_report(y_test, y_pred))

# Salva o modelo treinado em um arquivo
joblib.dump(clf, './model/trained_model.pkl')
joblib.dump(tfidf_vectorizer, './model/tfidf_vectorizer.pkl')
```

Os principais passos realizados no código são descritos a seguir.

1. **Importação de Bibliotecas:** Importa as bibliotecas necessárias para o código, incluindo Pandas para manipulação de dados, re para expressões regulares, scikit-learn para vetorização TF-IDF, divisão de dados, construção do modelo de Regressão Logística e métricas de avaliação, e joblib para salvar o modelo treinado e o vetorizador TF-IDF.

2. **Carregamento do Conjunto de Dados:** Carrega um conjunto de dados do arquivo CSV "SOCIOS_02.mixed.csv" usando a biblioteca Pandas. O conjunto de dados contém duas colunas, 'Data' e 'Tag'.

3. **Divisão do Conjunto de Dados:** Divide o conjunto de dados em conjuntos de treinamento e teste usando a função 'train_test_split' do scikit-learn. Os dados de entrada são da coluna 'Data', e os rótulos são da coluna 'Tag'. O conjunto de teste é definido como 20% do conjunto de dados total, e o valor de semente (random_state) é definido como 42 para garantir reprodutibilidade.

4. **Extração de Recursos TF-IDF:** Usa a classe 'TfidfVectorizer' do scikit-learn para converter o texto em recursos numéricos usando a técnica TF-IDF (Term Frequency-Inverse Document Frequency). Duas transformações TF-IDF são realizadas, uma para o conjunto de treinamento ('X_train_tfidf') e outra para o conjunto de teste ('X_test_tfidf'). Apenas as 1000 principais *features* são mantidas, definidas pelo parâmetro 'max_features'.

5. **Treinamento do Modelo de Regressão Logística:** Cria um modelo de Regressão Logística (‘clf‘) usando a classe ‘LogisticRegression‘ do scikit-learn e o treina com os dados de treinamento (‘X_train_tfidf‘ e ‘y_train‘) usando o método ‘fit‘.

6. **Avaliação do Modelo:** Realiza a previsão (‘y_pred‘) do conjunto de teste (‘X_test_tfidf‘) usando o modelo treinado e imprime um relatório de classificação (‘classification_report‘) que exibe métricas de avaliação, como precisão, recall e F1-score para cada classe no conjunto de teste.

7. **Salvando o Modelo Treinado:** Salva o modelo de Regressão Logística treinado (‘clf‘) em um arquivo chamado ’trained_model.pkl’ e também o vetorizador TF-IDF (‘tfidf_vectorizer‘) em um arquivo chamado ’tfidf_vectorizer.pkl’ usando a biblioteca joblib. Esses arquivos podem ser usados posteriormente para fazer previsões em novos dados sem a necessidade de treinar o modelo novamente.

6.5 Resultados do experimento

O resultado do primeiro experimento foi parcialmente satisfatório, pois considerando a característica do *dataset* utilizado, houveram poucas variações de formatos em campos do tipo documento como CPF e CNPJ. No caso do campo CPF, utilizou-se somente o formato com pontuação (968.573.055-52), e no CNPJ a base de origem tinha dados sem pontuação (12856119000163).

	precision	recall	f1-score	support
cnpj	0.85	1.00	0.92	8969
cpf	1.00	1.00	1.00	10038
nome	1.00	0.84	0.91	9930
estado	1.00	1.00	1.00	10051
accuracy			0.96	38988
macro avg	0.96	0.96	0.96	38988
weighted avg	0.97	0.96	0.96	38988

Figura 13 – Resultados Regressão Logistica

Na Figura 13, observa-se que cada um dos tipos de campo verificados demonstrou bons resultados de precisão, recall e f1-score. Em resumo, o resultado exibido nesta figura fornece uma análise detalhada de quão bem o modelo pode identificar cada classe (CNPJ, CPF, Nome e Estado), bem como uma avaliação geral de seu desempenho. De forma geral, o modelo apresentou alta precisão, recall e F1-score, indicando um bom desempenho, especialmente para as classes 1 e 3 (CPF e Nome).

Em um cenário de validação manual dos resultados, verificou-se bons resultados todas as vezes que foi mantido o formato do dado original utilizado no treinamento. Na Figura 14, o algoritmo identificou corretamente o meu nome (que não estava na base original, como sendo um dado do tipo 2 (nome). O mesmo comprovou-se em um teste com um estado, conforme Figura 15.

```
import joblib

# Load the trained model and TF-IDF vectorizer
loaded_model = joblib.load('trained_model.pkl')
loaded_tfidf_vectorizer = joblib.load('tfidf_vectorizer.pkl')

# Example of how to use the loaded model for prediction
new_text = "Felipe Casali Silva"
new_text_tfidf = loaded_tfidf_vectorizer.transform([new_text])
predicted_label = loaded_model.predict(new_text_tfidf)

print("Predicted Label:", predicted_label[0])
```

```
Predicted Label: 2
```

Figura 14 – Validação de string com formato de nome

```
new_text = "SP"
new_text_tfidf = loaded_tfidf_vectorizer.transform([new_text])
predicted_label = loaded_model.predict(new_text_tfidf)

print("Predicted Label:", predicted_label[0])
```

```
Predicted Label: 3
```

Figura 15 – Validação de string com formato de estado

Contudo, no caso dos campos de tipo de documento, conforme Figura 16, verificou-se que o algoritmo não conseguiu identificar um CNPJ com formato diferente do que foi fornecido no treinamento. Ao inserir pontos e traço no CNPJ, o algoritmo apontou erroneamente o dado como sendo do tipo 1 (CPF).

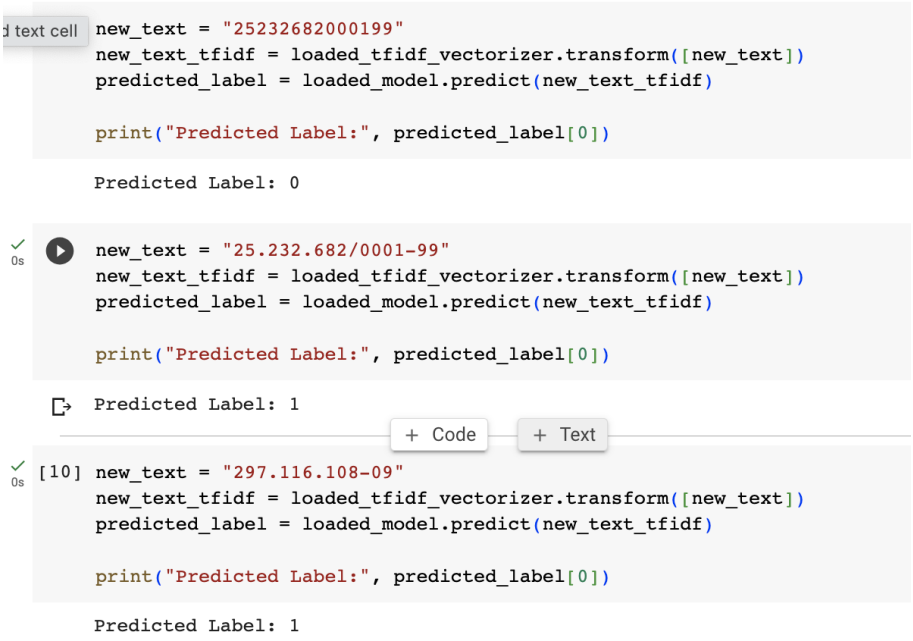

Figura 16 – Validação de string com formato de documentos

7 CONCLUSÃO

Neste capítulo é feita a conclusão deste trabalho de conclusão de curso, que tem como objetivo investigar o tema proteção dos dados na era da IA. Na seção 7.1 são destacadas as contribuições do trabalho. Na seção 7.2 são resumidas as dificuldades encontradas. Na seção 7.3 são listadas propostas para trabalhos futuros.

7.1 Contribuições

Apesar dos quase 20 anos de existência, a IA ganhou destaque nos últimos meses, principalmente por conta da popularização de ferramentas como o ChatGPT. Essa popularização aumentou as discussões sobre privacidade dos dados. Como resultado, hoje existem comissões em nível mundial discutindo como proteger a sociedade de possíveis danos causados pela tecnologia.

Entretanto, ainda existe um longo caminho a ser percorrido no tema de proteção de dados para projetos de IA. Este trabalho de conclusão de curso tem como objetivo investigar esse tema considerando aspectos teóricos e práticos.

Do ponto de vista de aspectos teóricos, o trabalho introduz as seguintes contribuições:

- Ilustra diferentes exemplos de fontes de dados com o objetivo de destacar que essas fontes possuem dados sensíveis e que devem ser manipulados de forma apropriada para garantir proteção.
- Realiza um delineamento do ciclo de vida dos dados, desde sua origem nas fontes de dados até o seu uso em algoritmos de IA.
- Sumariza os principais aspectos relacionados às leis e regulamentações para a proteção dos dados.
- Identifica e discute técnicas e ferramentas voltadas à proteção de dados sensíveis.
- Define sete estratégias que um cientista de dados deve empregar para ajudar na proteção dos dados em projetos de IA.

Utilizando como base a premissa de que IA pode contribuir para a proteção de dados sensíveis, o trabalho introduz a seguinte contribuição do ponto de vista prático:

- Desenvolve um modelo de IA baseado em regressão logística e expressões regulares e o aplica sobre dados sensíveis da LGPD.

7.2 Dificuldades Encontradas

Durante o desenvolvimento do trabalho, foram encontradas as dificuldades descritas a seguir. A primeira dificuldade refere ao fato de que são poucas as menções à proteção de dados em trabalhos oriundos de cursos de Tecnologia, Computação e Matemática. Conforme discutido na seção 1.2, a maioria dos trabalhos acadêmicos que discorrem sobre proteção de dados têm como origem cursos de Direito.

A segunda dificuldade refere-se ao desafio de sumarizar conteúdos oriundos de ambientes tão distintos, nos quais de um lado o objetivo é cumprir a legislação e do outro o objetivo é ter acesso à maior quantidade de informações que possa facilitar a tomada correta de decisões em ambientes de negócios.

Em terceiro lugar, ficou evidente que o nível de complexidade para identificação de números de documentos como CPF e CNPJ é alta, pois cada empresa pode armazenar esses dados em um formato diferente. Como não há uma padronização na forma como esse tipo de dado deve ser armazenado, é comum encontrar bases de dados que armazenam números de documentos com pontos e traços, e outras sem. Em situações extremas, os dois formatos podem ser encontrados em uma única coluna.

7.3 Trabalhos Futuros

Como trabalhos futuros, pode-se citar:

- A realização de um ajuste fino no modelo de IA desenvolvido, de forma que os resultados possam ser mais refinados.
- A utilização de outros modelos de IA e a posterior comparação dos resultados obtidos com os resultados descrito neste trabalho.
- A geração de *datasets* com dados reais, o que é fundamental para que o treinamento dos modelos de IA apresente resultados satisfatórios.
- O uso de fontes de dados de diferentes setores econômicos, a fim de garantir que o treinamento dos modelos de IA possa abranger ao máximo os campos sensíveis, estejam eles no contexto da LGPD ou não.
- O desenvolvimento de ferramentas capazes de identificar riscos de exposição por meio do cruzamento de dados de diferentes fontes.

REFERÊNCIAS

AISSI, M. E. M. E. *et al.* Data lake versus data warehouse architecture: A comparative study. *In*: BENNANI, S. *et al.* (ed.). **WITS 2020**. Singapore: Springer Singapore, 2022. p. 201–210. ISBN 978-981-33-6893-4.

ASHTON, K. *et al.* That 'internet of things' thing. **RFID journal**, Hauppauge, New York, v. 22, n. 7, p. 97–114, 2009.

CARROLL, R. J.; PEDERSON, S. On robustness in the logistic regression model. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 55, n. 3, p. 693–706, jul. 1993. Disponível em: https://doi.org/10.1111/j.2517-6161.1993.tb01934.x.

DIOUF, P. S.; BOLY, A.; NDIAYE, S. Variety of data in the etl processes in the cloud: State of the art. *In*: **2018 IEEE International Conference on Innovative Research and Development (ICIRD)**. [*S.l.: s.n.*], 2018. p. 1–5.

FANG, H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. *In*: **2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)**. [*S.l.: s.n.*], 2015. p. 820–824.

FORNASIER, M. D. O.; SPINATO, T. P.; RIBEIRO, F. L. Ransomware e cibersegurança: a informação ameaçada por ataques a dados. **Rev. Thesis Juris**, University Nove de Julho, v. 9, n. 1, p. 208–236, jun. 2020.

HARRISON, M. **Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python**. [*S.l.: s.n.*]: Novatec Editora, 2019.

LI, J. *et al.* Feature selection: A data perspective. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 50, n. 6, dec 2017. ISSN 0360-0300. Disponível em: https://doi.org/10.1145/3136625.

LUGATI, L. N.; ALMEIDA, J. E. d. Da evolução das legislações sobre proteção de dados: a necessidade de reavaliação do papel do consentimento como garantidor da autodeterminação informativa. **Revista de Direito**, v. 12, n. 02, p. 01–33, ago. 2020. Disponível em: https://periodicos.ufv.br/revistadir/article/view/10597.

LUGATI, L. N.; ALMEIDA, J. E. de. Da evolução das legislações sobre proteção de dados: a necessidade de reavaliação do papel do consentimento como garantidor da autodeterminação informativa. **Revista de Direito**, Revista de Direito, v. 12, n. 02, p. 01–33, ago. 2020. Disponível em: https://doi.org/10.32361/2020120210597.

O'KANE, P.; SEZER, S.; CARLIN, D. Evolution of ransomware. **IET Networks**, v. 7, n. 5, p. 321–327, 2018. Disponível em: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-net.2017.0207.

PARLIAMENT, E. **EU Legislation in Progress - Artificial intelligence act**. 2022. Disponível em: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

QUEIROZ, R. C. Z. **A proteção de dados pessoais: a LGPD e a disciplina jurídica do encarregado de proteção de dados pessoais**. 2023. Tese (Doutorado), 2023.

SOUSA, L. R. **Analytics: critical success factors on implementation in organizations**. 2018. Tese (Doutorado), 2018.

WREMBEL, R. Still open problems in data warehouse and data lake research: extended abstract. *In*: **2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS)**. [*S.l.: s.n.*], 2021. p. 01–03.