**ESTUDO DE CASO PARA MÉTODOS DE APLICAÇÃO DA LEI GERAL DE PROTEÇÃO DE DADOS**

Juan Cardoso da Silva, Ronaldo Celso Messias Correia

Departamento de Matemática e Computação (DMC)

Universidade Estadual Paulista “Júlio Mesquita Filho” (UNESP)

Faculdade de Ciências e Tecnologia (FCT)

Presidente Prudente - SP, Brasil

[**juan.c.silva@unesp.br**](mailto:juan.c.silva@unesp.br)**,** [**ronaldo.correia@unesp.br**](mailto:ronaldo.correia@unesp.br)

**Resumo – Com a ampliação das leis protetoras de dados digitais privados ao redor do mundo como G.D.P.R e a L.G.P.D. - Surgiu um espaço não explorado e sem atenção até as décadas recentes, a interseção das ciências humanas com as ciências exatas, a aplicação de Inteligências Artificiais para a manipulação desses dados dentro do contexto da área de direito para de alguma forma facilitar o cumprimento da lei em garantir a persistência dos dados e sua proteção. Neste trabalho será apresentado um processo de análise e identificação de dados privados por meio de diversos modelos de máquinas criados, com objetivo de analisar o potencial dessas máquinas ao identificar os dados como sua capacidade de recomendar encriptações como um método de proteção de tais dados.**

**Abstract – The**

***Palavras Chaves –* Lei Geral de Proteção de dados (LGPD); Proteção de dados; Métodos de aplicação da LGPD, Manipulação e minimização de dados; Machine Learning; Inteligencia Artificial;**

# INTRODUÇÃO

Com a evolução da tecnologia focada em conectar computadores se tornou disponível para as pessoas nos anos 90, surgiu a World Wide Web e futuramente, nomeado Web 1.0 – Sendo um sucesso mundial e fornecendo diversos sites focados em encontrar informações espalhadas na internet, geralmente contendo sites estáticos e sem muitos detalhes e funcionalidades, também, essa versão da web fornecia atualizações para sistemas e softwares, tais como o rudimentar Windows Update do Windows 95 e sites de terceiros dedicados para jogos como Quakeworld (hoje em dia continuado por fãs, o web site original está fechado), nessa época a internet vivia sem rumo sobre sua funcionalidade concreta no futuro, até que em 1999 surgiu o primeiro uso do termo para sua segunda iteração, a Web 2.0

A Web 2.0 foi pensada na interatividade entre pessoas na internet além do acesso e busca de informações, com isso diversos sites novos começaram a nascer focados em fornecer serviços como os de construções de *Wikis*, Aplicações Web, *Mashups*, Redes Sociais, Eventos Colaborativos (caridades e doações para determinados motivos tais quais crowdsourcing e crowdfunding, ambos para financiar tipos de projetos, tais como *Patreon* e *Kickstarter*), Conteúdo gerado de usuários (*Youtube*), Blogs, Streaming Curadoria Social (*Reddit* e *Instagram*). Com esse serviço disponível agora em mãos dos usuários e popularizado ao ponto de dados sensíveis como cartões de créditos, emails e outras informações pessoais circulando, era inevitável a chegada de agentes indesejáveis atacando a rede em busca de acesso a esses dados, buscando lucro em tentativas de se apropriar desses dados.

Os ataques desses agentes eventualmente resultaram na criação de leis progenitoras, focadas em tratar ataques relacionados a privacidades, como a Lei de Combate a Crimes Cibernéticos (12.737/2012) criada para tratar o caso da Carolina Dieckmann. Em 2016 um evento no Reino Unido onde o Facebook indevidamente estava utilizando dados privados de alunos em uma universidade, acabou resultando na *General Data Protection Regulation* (GDPR), dois anos depois, o congresso brasileiro criou sua resposta em forma de lei, utilizando a lei 12.737/2012 e bases da GDPR para criar a Lei Geral de Proteção de dados (LGPD). Após sua implementação em maio de 2021, diversos métodos de implementação da lei foram adotados para respeita-la, de novas opções de gerenciamentos de cookies até como os dados são manipulados no em sistemas.

Neste artigo será apresentado como a LGPD (a aplicação dos incisos I, VII e VIII) podem serem utilizados em conjunto de modelos de machine learning e comparar os resultados dos modelos a apresentados, suas limitações quanto a capacidade de categorizar os dados de uma maneira binária e assegurar os incisos mencionados (serão apresentados na sessão de Fundamento Teórico) sejam respeitados.

# FUNDAMENTAÇÃO TEÓRICA

Esta seção é dedicada a discussão sobre os conceitos básicos relacionados aos temas da LGPD, Machine Learning e encriptação.

# LEI GERAL DE PROTEÇÃO DE DADOS

Criada em 2018 e entrado em vigor em 2021, a Lei Geral de Proteção de Dados 13.709/2018 (abreviando como LGPD) possui o objetivo de garantir os direitos de privacidade e proteção de dados sensíveis de usuários e o livre desenvolvimento da personalidade pessoal natural – a lei atua também como um mecanismo de segurança judiciária, promovendo padronizações, regulamentos, práticas de proteção de dados pessoais/sensíveis e a punição por multas e até prisões, caso ocorra não realização da lei.

Pela lei, dados sensíveis/privados são informações relacionadas diretamente a intimidade pessoal de um indivíduo, opiniões políticas, convicções religiosas, filiação a organizações religiosas/políticas/filosóficas, etnia, dados relacionados a saúde, dados genéticos (como biometria e afins), CPF e todo outro tipo de informação do qual possa ser vinculada a vida íntima de uma pessoa.

A lei garante ao cidadão a capacidade de deleção dos dados sensíveis/privados, revogar consentimento do uso de tais dados, transferir os dados para outro serviço similar ao serviço atual, acesso aos dados para atualiza-los e garantir a consistência dos dados no meio digital.

Além disso a lei [5], também dispõe significados para como esses dados podem ser manipulados, pela lei, o tratamento de dados é entendido como qualquer tipo de operação onde os dados pessoais recebem interação por parte de um profissional, tais como acessar um dado, alterar, deletar, produzir, coletar, classificar, acessar, reproduzir, copiar, processar, armazenar e quais queres outras manipulações possíveis com os dados.

A lei também determina a existência de um operador e um controlador, operadores são pessoas das quais possam tomar decisões finais a respeito de como o dado vai ser manipulado e os operadores realizam as manipulações, ambos podem ser pessoas naturais ou jurídicas, sendo elas do âmbito público ou privado [5].

Pelas considerações do Art. 6 [5], o tratamento de dados pessoais deve seguir alguns princípios, para este projeto, os princípios considerados para atingir o objetivo procurado são transparência, adequação, finalidade e responsabilização.

Os princípios seguem como [5]:

1. **Finalidade:** realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;
2. **Adequação:** compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;
3. **Necessidade:** limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;
4. **Livre Acesso:** garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;
5. **Qualidade de dados:** garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;
6. **Transparência:** garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;
7. **Segurança:** utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;
8. **Prevenção:** adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;
9. **Não discriminação:** impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;
10. **Responsabilização:** demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas.

# MACHINE LEARNING

*Machine Learning*, uma categoria pertencente à área de IA tem como objetivo “treinar” máquinas com intenção de resolver problemas, automatizar tarefas e identificar padrões.

Com as capacidades modernas da *Machine Learning* aumentando, a área deixou de apenas ser estruturação e dedução de dados, evoluindo para entender padrões indefinidos usando técnicas de treinamento do qual cada padrão aprendido utilizado, pode ser utilizado para analisar, outros padrões indefinidos, compreender dados e automatizar tarefas. Hoje me dia, com a capacidade computacional da nuvem e a abrangente quantidade de dados disponíveis, ou até mesmo gerados por outras ML para treinar ML, demonstrando a capacidade de ser aplicada em diversos cenários diferentes para chegar a um resultado esperado/procurado, ajudando em várias áreas de atuações nas profissões, com as maiores contribuições da ML sendo para as áreas da saúde e áreas de processamento de imagens. Este trabalho foca em trazer essas ideias de aplicações de ML para área de direito aplicada, mais especificamente na aplicação de LGPD.

Utilizando *Deep Learning* podemos usar a ML para aprender a não só reconhecer e identificar padrões, como utilizar para realizar avaliações de dados, seguindo algumas categorias de avaliações.

# TENSORFLOW

*TensorFlow*, é um framework de algoritmos de *Deep* *Learning* e *Machine* *Learning*, desenvolvido pela *Google* *Brains* e tornado *open-source* em 2015, sendo uma biblioteca aberta utilizando computação numérica em larga escala.

O *TensorFlow* cria grafos/estruturas permitindo o fluxo destes sejam processados, cada vértice do grafo possui a representatividade de operações matemáticas e cada aresta, um array multidimensional, chamado de *tensor*.

A biblioteca permite funcionar localmente, consumindo GPU e CPU da máquina do programador ou utilizando uma máquina na nuvem, com um maior poder computacional, neste caso, utilizando uma TPU resultando em maior poder computacional. No desenvolvimento do projeto no entando, foi utilizado GPU em conjunto da CPU localmente em um tensorflow instalado localmente na máquina e rodando dentro de um ambiente curado no jupyter notebook onde foi realizados tratamento de dados, criação de modelos e testes.

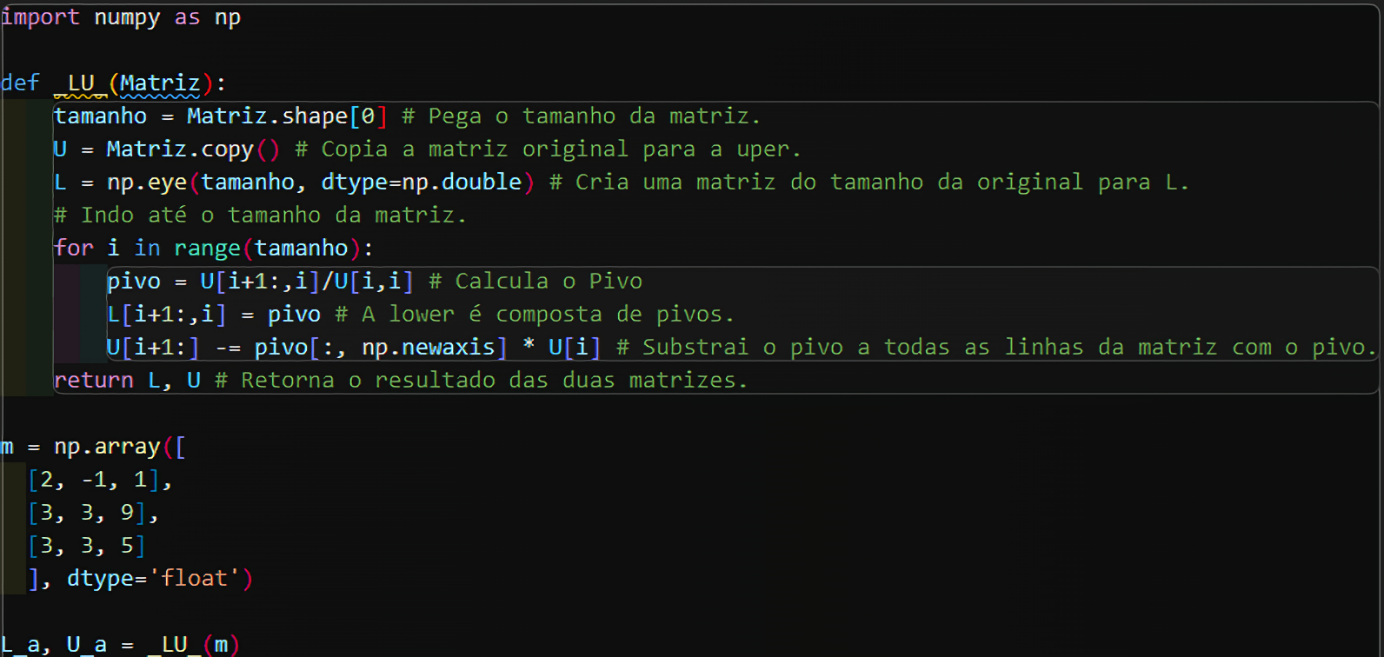
# JUPYTER NOTEBOOK

Jupyter notebook é um ambiente para executar e projetar códigos na web utilizando python como sua linguagem de programação, cada notebook é dividido em seções acompanhadas com áreas de textos contendo markdown text para anotações e títulos. Cada seção dentro de um notebook (chamada de documentos) possui um input e um output, servindo como um espaço de log para cada documento e permitir verificar resultados anteriores como permitir uma modularidade maior para corrigir e escrever código.

# MATPLOTLIB, NUMPY E PANDSA

Numpy é uma biblioteca de Python criada utilizando vetorização invés de loops, arrays e indexação, utilizando código de máquina pré compilado em C baseada no objeto *narray* da biblioteca para realizar as operações com velocidade e consistência, pois todas as operações necessárias estão pré-compiladas e armazenadas pronto para serem utilizadas.

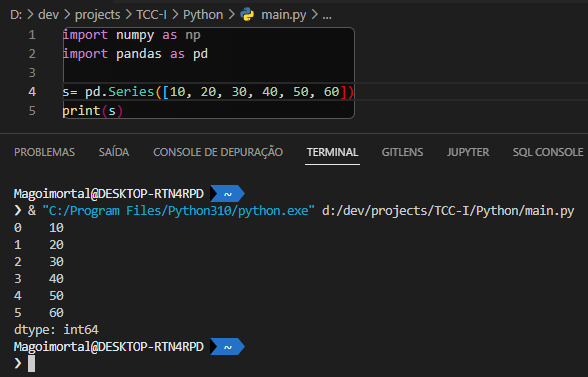
Figura 1 Exemplo de operação com matrizes utilizando NumPy.



Fonte: Elaborado pelo Autor

Pandas permite o cientista poder manipular os dados sem necessitar preocupar com tratamento das estruturas, complexidade e tempo de velocidade dentro de um algoritmo, facilitando a operações/manipulações com dados, estruturando-os para uso em conjunto de outras aplicações, como aprendizado de máquina, visualização de dados ou até mesmo Cálculo Numérico.

Figura 2 Exemplo de uma estrutura de série utilizando os pandas.



Fonte: Elaborado pelo autor.

Matplotlib é uma API implementada em Python para visualização de dados em forma gráfica, permitindo visualização de dados conforme a necessidade do programador ao realizar testes, experimentos, pesquisas e trabalhar com resultados obtidos de projetos.

# CRIPTOGRAFIA

RSA é um algoritmo de criptografia desenvolvido em 1976 logo depois do primeiro algoritmo de compartilhamento de chaves chamado Diffie-Hellman apresentado. O RSA permite a encriptação de dados e a desencriptarão dos mesmos utilizando uma chave pública, privada e um números primos para geração dessas chaves.

Primeiro selecionamos um valor P e Q diferentes e quando multiplicado sejam maiores que 400, depois calculamos o valor N com multiplicação de P por Q, em seguida calculamos o euler totiente:

Depois seleciona um inteiro E que seja co-primo do nosso euler totiente e procedemos a calculamos nosso D:

Sabendo nosso D e E podemos criar as chaves públicas e privadas e fazer as encriptações de ida e volta conforme a necessidade:

Processo de ida:

Processo de volta

# FAKER API

Faker API é uma API de geração de dados utilizados para construções de projetos, testes unitários e criação de backends – cada dado gerado pelo fake é razoável em relação a um dado “real”. A API foi escrita primeiramente em PERL e depois portada para outras linguagens de programações, a API permite também que a formação desses dados seja baseada no país determinado pelo programador.

# TRABALHOS RELACIONADOS

A IBM realizou um trabalho onde foi desenvolvido um sistema automatizado na nuvem para detecção de irregularidades em uma base de dados, chamado de QRadar [3]. Este sistema realiza varreduras para confirmar se existiu casos onde os dados sensíveis foram deletados com sucesso da base, se existir vestígios ou uma possibilidade de a deleção não ocorrer com sucesso, um trigger é ativado para realizar outra varredura mais profunda, os modelos apresentados anteriormente, realizar encriptação dos dados para estar de acordo com o inciso de permanência dos dados, já que caso necessário por investigação, eles precisam estar disponíveis, entretanto, os modelos não possui a capacidade de varreduras profundas como o modelo da IBM.

Neste trabalho de pesquisa [4], também da IBM, o modelo de ML utilizado por pelos pesquisadores usa escolhas binárias para os dados coletados serem minimizados de acordo com a GDPR, no caso, estão a utilizar dado de um hospital para escolher quais dados podem ser deletados baseado nos parâmetros escolhidos. Comparado aos modelos, apesar de um resultado mais condizente com a LGPD, existe uma porcentagem pequena de valores não encriptados como falsos negativos, onde neste projeto, pelo escopo mais simples, foi possível atingir o corte dos dados necessários dos salários para aparecer na dashboard de dados.

# METODOLOGIAS

A utilização do Faker API é um dos primeiros pontos a ser discutidos na aplicação do projeto, por ser uma API de geração de dados razoáveis porém falsos – os dados gerados não são relacionados a pessoas reais apesar de terem todas características de dados sensíveis (relacionados a pessoas reais e sua privacidade, como definido na LGPD), isso permite a experimentação mais aberta e sem precisar utilizar empresas fornecedoras de dados privados sem gastar recursos, como arriscar vestígios de dados sensíveis nos melhores pesos dos modelos *de machine leraning.*

Utilizando RSA é possível pegar os dados fornecidos pelo Faker API e encripta-los de uma maneira onde é possível realizar o retorno dos dados para serem visualizados como a lei LGPD pede, utilizando-a em conjunto com os modelos, cada dado avaliado será tratado como condizente como dado sensível ou não, se for, será encriptado.

# LIMITAÇÕES

BBB

# REFERÊNCIA BIBLIOGRÁFICA

**[1] Lei 'Carolina Dieckmann', que pune invasão de PCs, entra em vigor.** G1 Globo, São Paulo, dia 1, junho de 2013. Disponível em: <https://g1.globo.com/tecnologia/noticia/2013/04/lei-carolina-dieckmann-que-pune-invasao-de-pcs-passa-valer-amanha.html> Acesso em: 27, agosto de 2022

**[2]** Leal da Silva, Julia; **Tomada de Decisão Automatizada e Controle pela LGPD.** IAPD, 20, janeiro de 2021. Disponível em: <https://iapd.org.br/decisao-automatizada-lgpd-direito-aexplicacao/> Acesso em: 24, agosto de 2022.

**[3]** Fortunato, Caroline, **Using QRadar for LGDP**, 19, julho de 2019, Disponível em:<https://www.proof.com.br/wp-content/uploads/2019/08/Using-QRadar-for-LGPD.pdf>

**[4]** Goldsteen, Abigail; Ezov Gilad; Shmelkin, Ron; Moffie, Micha; Farkash, Ariel; **Data minimization for GDPR compliance in machine learning models**, p. 1-15, 2021, Disponível em: <<https://link.springer.com/article/10.1007/s43681-021-00095-8>>

**[5]** Roberto Fernandes Castilho, José; Andrade Gomes, Henrique; **Legislação Básica de Direito da Informática**, 2° edição reformulada e atualizada, São Paulo, Editora Pillares.

**[6] General Data Protection Regulation GDPR:** OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018. Disponível em: <<https://gdpr-info.eu/>> Acesso em: 27, agosto de 2022

**[7]** Stallings, William; Brown, Lawrie; **Computer Security: Principles and Practice Second Edition**, 2° edição reformulada e atualizada, São Paulo, Pearson; 2ª edição.

**[8]** Mota, José; **Da Web 2.0 ao E-Learning 2.0: Aprender na rede,** p.1-199, 2009. Disponível em: <https://repositorioaberto.uab.pt/handle/10400.2/1381>

**[9]** Filipe Lima Rapôso, Cláudio; Melo de Lima, Haniel; Ferreira de Oliveira Junior, Waldecy; Aragão Ferreira Silva, Paola; Elaine de Souza Barros, Elaine; **LGPD - LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS EM TECNOLOGIA DA INFORMAÇÃO: Revisão Sistemática**, p. 1-10. 2019. Disponível em:<https://revistas.cesmac.edu.br/index.php/administracao/article/view/1035>

**[10]** Spadaccini de Teffé, Chiara; Viola, Mario; **Tratamento de dados pessoais na LGPD: estudo sobre as bases legais**, p. 1-38, 2020. Disponível em: <https://civilistica.emnuvens.com.br/redc/article/view/510>

**[11]** Tankard, Colin; **What the GDPR means for business**, p. 1-8, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1353485816300563?casa\_token=pjOHlq5iYsAAAAA:dhlGukGxSanwjnBY9aPtw36O1CslOJG1wZ7wTEryTJHi5QtFVM2G6kc8CwuemdVrUDRl2tRqJoA>

**[12]** Goldsteen, Abigail; Ezov Gilad; Shmelkin, Ron; Moffie, Micha; Farkash, Ariel; **Data minimization for GDPR compliance in machine learning models**, p. 1-15, 2021. Disponível em: <https://link.springer.com/article/10.1007/s43681-021-00095-8>

**[13]** Shanmugam, Divya; Shabanian, Samira; Diaz, Fernando; Finck, Michèle, Biega, Asia; **Learning to Limit Data Collectionvia Scaling laws: A computional Interpolation for the Legal Principle of Data Minimization**, p. 1-11, 2022. Disponível em: < https://arxiv.org/abs/2107.08096 >

**[14]** D. R. Ignatius Moses Setiadi, A. Faishal Najib, E. H. Rachmawanto, C. Atika Sari, K. Sarker and N. Rijati, "**A Comparative Study MD5 and SHA1 Algorithms to Encrypt REST API Authentication on Mobile-based Application**," 2019 International Conference on Information and Communications Technology (ICOIACT), 2019, pp. 206-211, doi: 10.1109/ICOIACT46704.2019.8938570. Disponível em: <https://ieeexplore.ieee.org/document/8938570>

**[15]** Boutaba, Raouf; Mohammad A., Salahuddin; Limam, Noura; Ayoubi, Sara; Shahriar, Nashid; Estrada-Solano Felipe; Caicedo M. Oscar; **A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities**, p. 1-99, 2018. Disponível em: <https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0087-2#Sec2>

**[16]** Fatih, Ertam; Galip, Aydin; **Data classification with deep learning using TensorFlow**, p. 1-4, 2017. Disponível em: <https://ieeexplore.ieee.org/document/8093521>

**[17]** Nguyen Quang-Hung; Hieu Doan; Nam Thoai; **Performance Evaluation of Distributed Training in TensorFlow 2**, p. 1-5, 2020. Disponível em: < https://ieeexplore.ieee.org/document/9353085 >

**[18]** PYTHON SOFTWARE FOUNDATION. **Python Language** Site: Documentation, 2022. Disponível em: <https://www.python.org/doc/> Acesso em: 24 de agosto de 2022.

**[19]** PANDAS **Pandas DataFrame** Site: About, 2022. Página sobre nós. Disponível em: <https://pandas.pydata.org/about/> Acesso em: 24 de agosto de 2022.

**[20]** PANDAS **Pandas DataFrame** Site: Documentation, 2022. Disponível em: <https://pandas.pydata.org/about/> Acesso em: 24 de agosto de 2022.

**[21]** PANDAS **10 minutes to pandas** Site: User Guide, 2022. Disponível em: <https://pandas.pydata.org/docs/user\_guide/10min.html#viewing-dataAcesso em: 24 de agosto de 2022.

**[22]** NUMPY PROJECT AND COMMUNITY. **What is Numpy?** Site: Documentation, 2022. Página de documentação. Disponível em:

<https://numpy.org/doc/stable/user/whatisnumpy.html> Acesso em: 24 de agosto de 2022.

**[23]** J. D. Hunter, "**Matplotlib: A 2D Graphics Environment**," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007, doi: 10.1109/MCSE.2007.55. Disponível em:<https://ieeexplore.ieee.org/document/4160265>

**[24]** MDN Web Docs, **Mozilla Developer Network Web Docs**, 2022. Disponível em: <https://developer.mozilla.org/pt-BR/>

**[25]** CLOUD GOOGLE. **Considerações sobre dados confidenciais em conjuntos de dados de aprendizado de máquina.** Site: Documentação, 2022. Disponível em: <<https://cloud.google.com/architecture/sensitive-data-and-ml-datasets>> Acesso em: 27, agosto de 2022.

**[26]** D Rachmawati, J T Tarigan and A B C Ginting; **A comparative study of Message Digest 5(MD5) and SHA256 algorithm**, p. 1-7, 2018. Disponível em <<https://iopscience.iop.org/article/10.1088/1742-6596/978/1/012116/meta>>

**[27]** Song, Congzheng; Ristenpart, Thomas; Shmatikov, Vitaly; **Machine Learning Models that Remember Too Much**, p. 1-15, 22, setembro de 2017. Disponível em: <<https://arxiv.org/pdf/1709.07886.pdf>>

**[28]** Vieira Souza, Iuri; **Aplicações de software desenvolvidas no contexto da Inteligência Artificial (IA), Machine Learning e Big  
Data e o direito dos cidadãos de acordo com a Lei  
Geral de Proteção de Dados (LGPD)**, p. 1-83, 2021. Disponível em: < https://bdm.unb.br/bitstream/10483/30275/1/2021\_IuriSousaVieira\_tcc.pdf>

**[29]** Goldsteen, Abigail; Ezov Gilad; Shmelkin, Ron; Moffie, Micha; Farkash, Ariel; **Data minimization for GDPR compliance in machine learning models**, p. 1-15, 2021, Disponível em: <<https://link.springer.com/article/10.1007/s43681-021-00095-8>>

[30] K. Pavani and P. Sriramya, "Enhancing Public Key Cryptography using RSA, RSA-CRT and N-Prime RSA with Multiple Keys," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1-6, doi: 10.1109/ICICV50876.2021.9388621. Disponível em: <https://ieeexplore.ieee.org/document/9388621>