

**FCT – Faculdade de Ciências e Tecnologia**

**DMC – Departamento de Matemática e Computação**

**Bacharelado em Ciência da Computação**

Trabalho de Conclusão de Curso

(Modalidade Trabalho Acadêmico)

REVISÃO DE LITERATURA

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

**Autor:** Juan Cardoso da Silva

**Orientador:** Prof. Ronaldo Celso Messias Correia

**Presidente Prudente**

**2022**

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

Revisão de Literatura apresentada ao Curso de Ciência da Computação do Departamento da Matemática e Computação da Universidade Estadual Paulista – Unesp, campus de Presidente Prudente, como requisito obrigatório para aprovação na disciplina Trabalho de Conclusão de Curso I, ministrada pelo Prof. Dr. Celso Olivete Júnior.

**Orientador:** Prof. Ronaldo Celso Messias Correia

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

FACULDADE DE CIÊNCIA E TECNOLOGIA

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**Presidente Prudente**

**2022**

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

**Autor:**  Juan Cardoso da Silva

**Orientador:** Prof. O Dr. [Ronaldo Celso Messias Correia](mailto:ronaldo.correia@unesp.br)

**Coorientador:** Prof. Dr.

**Responsável:** Prof. Dr. Celso Olivete Júnior

**Disciplina:** Trabalho de Conclusão de Curso I

**Curso:** Bacharelado em Ciência da Computação

**Instituição:** Universidade Estadual Paulista

        Faculdade de Ciência e Tecnologia

        Departamento de Matemática e Computação

Presidente Prudente, 28 de agosto de 2022.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Juan Cardoso da Silva (Aluno)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. O Dr. Ronaldo Celso Messias Correia (Orientador)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. A Dr. (Coorientador)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. O Dr. Celso Olivete Júnior

**LISTA DE ILUSTRAÇÕES**

[Figura 1 - Exemplo de operação com matrizes utilizando NumPy. 15](#_Toc112598868)

[Figura 2 - Exemplo de uma estrutura de série utilizando os pandas. 16](#_Toc112598869)

[Figura 3 - Comparação de erro entre as funções implementadas 17](#_Toc112598870)

[Figura 4 - Comparativo entre métodos de Euler criados e o método do NumPy. 17](#_Toc112598871)

[Figura 5 - Processo de formatação de mensagem 19](#_Toc112598872)

[Figura 6 - Processamento de mensagem SHA-512 20](#_Toc112598873)

**LISTA DE TABELAS**

**LISTA DE ABREVIATURAS E SIGLAS**

LGPD Lei Geral de Proteção de Dados

GDPR General Data Protection Regulation

ML Machine Learning

IA Inteligência Artificial

CPF Cadastro de Pessoas Físicas

SHA Security Hash Algorithm.

TPUTensorFlow Processing Unit

API Application Programming Interface

REST Representional State Transfer

**SUMÁRIO**

1. INTRODUÇÃO10
2. OBJETIVOS11
3. FUNDAMENTAÇÃO DA TEÓRICA12

3.1. Lei Geral de Proteção de Dados13

3.1.2. Disposições Preliminares13

3.1.3. Terminação de Dados13

3.2. Machine Learning14

3.2.1. TensorFlow14

3.2.2. NumPy15

3.2.3. Pandas16

3.2.4. Matplotlib16

3.2.5. Minimização de Dados17

3.2.6. Segurança da Informação18

3.2.7. Criptografia18

3.2.8. Encriptação com SHA-51218

3.2.8.1. Formatação de Entrada19

3.2.8.2. Inicialização de Buffer20

3.2.8.3. Processamento de Entrada20

3.2.8.4. Realização do digest21

3.2.9. Encriptação com RSA21

1. TRABALHOS RELACIONADOS22
2. CONCLUSÕES23
3. BIBLIOGRAFIAS24

**RESUMO**

Após a aprovação e implementação da Lei Geral de Proteção de Dados em 2018 e a entrada da lei em vigor em 2021, diversos logs de softwares e bases de dados tiveram um novo obstáculo: a proteção de dados de acordo com a lei e como melhor adequar os logs dos softwares ou base de dados à lei. A utilização de Machine Learning tem se tornado uma opção relativamente popular devido a capacidade de criar modelos para analisar as bases de dados ou logs de softwares procurando por implementações que não correspondem com o aprovado na lei. Este trabalho tem como objetivo criar um estudo de caso onde um modelo de Machine Learning será desenvolvido para filtrar dados em uma base de dados, utilizado métodos de classificação binária para julgar quais dados dentro da base devem ser tratados, utilizando criptografia para garantir a anonimidade dos dados sensíveis e descartando os dados irrelevantes na base de dados, sendo eles sensíveis ou não.

**Palavras-Chave:** Lei geral proteção de dados; LGPD projetos; Métodos de análise da LGPD; Manipulação e minimização de dados; Machine Learning;

**ABSTRACT**

After the approval and implementation of General Data Protection Law in 2018 and the law passing in 2021, a lot of logs from softwares, databases have gone through a new obstacle, the application of the protection law in accordance with it, better adequing the software or the database to the law. Using Machine Learning, it has been a popular option because of its capacity in creating models to analyze databases or software logs looking for datas that are not stored accordingly with the law. This work has the objective to create a case of study with a new made model of Machine Learning to filter these databases and logs, using binary classification to judge which data is within the law and which data needs to be processed to be within the law, using cryptography to anonymize sensitive data e discarding any irrelevant data from the database, regardless if they are sensitive or not.

**Keywords:** General Data Protection Law; LGPD; LGPD analysis Methods; Data manipulation and minimization; Machine Learning;

# INTRODUÇÃO

Com a popularização do acesso à internet, a web 1.0 se tornou um fenômeno mundial, conectando pessoas com chats, fóruns de diversos tipos (filmes, fofocas, jogos), primeiras redes sociais e os primeiros serviços online, mas não se tinha uma ideia do que exatamente a internet deveria se tornar.

Em 1999, o conceito protótipo de web 2.0 foi apresentado ao público da tecnologia com um artigo publicado na *Print Magazine*, em 2003, em uma sessão de brainstorming por Tim ‘O Reilly, visava transformar a internet em uma plataforma de serviços e acesso a softwares online, trazendo à tona, o uso dados pessoais como CPFs, número de cartões de créditos, e-mails, endereços físicos, números de telefones e outros dados sigilosos nesses serviços online, eventualmente, chegando no momento onde dados pessoais são utilizados nesses serviços para vender propaganda por meio de *Targeted advertising*, recomendações e dentre outros.

A exposição dos usuários na internet não só trouxe muitos benefícios em geral, como também abriu a porta para eventuais ataques na rede com objetivo de invadir não só o espaço privado da pessoa (tanto nos serviços online e nos computadores pessoais), como também para roubar dados pessoais para uso profano, como foi o caso da Carolina Dieckmann, que teve fotos intimas espalhadas na internet, neste caso os hackers, fizeram um *Ransom* das informações e fotos dela em troca de dinheiro, esse acontecimento foi a semente para que a Lei Geral de Proteção de Dados Pessoais fosse criada.

Na Europa, ocorreram diversos casos similares de vazamento de dados que resultaram na criação da *General Data Protection Regulation*, sendo um regulamento na União Europeia, que trata a privacidade online como um direito, sendo necessário proteger esta privacidade e evitar a exportação desses dados pessoais ao exterior. Com base nessa lei, a Lei Geral de Proteção de Dados Pessoais foi criada em 2018 e entrou em vigor em agosto de 2021.

# OBJETIVOS

No presente projeto, visa-se a utilização de uma framework para ML e projetar um modelo passível de realizar o processamento de dados sensíveis de uma base de dados, conseguindo indicar quais dados sensíveis precisam de atenção para entrar de acordo com o artigo 6 da LGDP, sendo esse modelo projetado capaz de processar e aprender quando um dado é sensível e está vulnerável, avisar sobre a vulnerabilidade e encriptar os dados de volta para a base de dados, caso o dado sensível não seja de uso da empresa, ele será descartado como parâmetro e será indicado para remoção da base de dados.

A criptografia utilizada será um algoritmo de criptografia do tipo *Hashing* do qual trabalha a entrada de um dado e retorna uma única hash do dado de entrada, sendo este o indicado a ser armazenado na base de dados, e não o dado de entrada.

A framework utilizada para o projeto será o TensorFlow em conjunto com linguagens como Python, utilizando bibliotecas como NumPy, Pandas com o TensorFlow, ou usando JavaScript diretamente com o TensorFlow em conjunto do WebGL.

Feito isso, será feito um estudo de como esse tipo de modelo de ML projetado se comporta e a viabilidade da implementação de um projeto deste para ajudar programadores e cientistas a estarem dentro da lei do artigo 6 da LGPD;

# FUNDAMENTAÇÃO TEÓRICA

**3.1. LEI GERAL DE PROTEÇÃO DE DADOS**

Após um escândalo do qual os dados da Cambrige University Analytics estavam sendo utilizados/minerados (processo aonde explora-se dados em busca de padrões em busca de achar um novo conjunto de dados relevantes), ilegalmente pelo Facebook e revendidos a terceiros, a união Europeia fez um esboço em 2016 e eventualmente a implementação da *General Data Protection Regulation* (GDPR) em 2018, focada em direitos individuais, proteção de dados pessoais, encriptação, gerenciamento de cookies, consentimento de uso de dados pessoais, segurança de dados e dentre outros mais.

Lei Geral de proteção de Dados (LGPD) é a nova lei brasileira inspirada na GDPR e construída em cima do Marco Civil da Internet, a lei é aplicada na área da computação focada na segurança de dados, proteção de dados, proteção à privacidade, criada para ser mais um pilar nas leis brasileiras para contribuir na criação/adaptação de ambientes de softwares se tornarem mais seguros, respeitando a privacidade do cidadão e garantindo a segurança dos seus dados sigilosos.

A base utilizada pela LGPD para o projeto é a proteção de dados pessoais em bases/bancos de dados em computadores, verificando se esses dados sensíveis estão de acordo com as normas da lei e se existe a necessidade de alteração.

**3.1.2. DISPOSIÇÕES PRELIMINARES**

A diretriz básica da LGPD pode ser entendida pelo Art. 2° da lei:

# o respeito à privacidade;

1. a autodeterminação informativa;
2. a liberdade de expressão, de informação, de comunicação e de opinião;
3. a inviolabilidade da intimidade, da honra e da imagem;
4. o desenvolvimento econômico e tecnológico e a inovação;
5. a livre iniciativa, livre concorrência e a defesa do consumidor; e
6. os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais.

A lei atribui os determinados significados para a esfera da intimidade dos dados pessoais, o dado pessoal é entendido como a informação pertencente o(a) um brasileiro(a) natural, esse dado pode ser dividido em dado pessoal sensível e anonimizado, este último é o dado de um indivíduo anônimo nos meios de comunicações disponíveis – o dado sensível é considerado a informação associada/pertencente diretamente ao indivíduo, seja ela CPF, data de nascimento, nome completo, pensamento político, genético e dentre outros – circulantes no meio de comunicação disponíveis.

Além disso a lei 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.

Pelas considerações do Art. 6, 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.

**3.1.3. TERMINAÇÃO DE DADOS PESSOAIS**

A configuração da lei determina, a terminação dos dados pessoais armazenados por terceiros a partir do contrato legal, em determinado momento entre usuário e terceiros seja cancelado (pelo usuário) ou terminado (fim de contrato), a exclusão de tal dados da base.

Existindo apenas algumas exceções que possam ocorrer na terminação:

1. órgãos de pesquisas que necessitem dos dados para estudo, mas estes devem anonimizar os dados para apresentação pública dentro do possível;
2. a lei obrigue por algum motivo, a utilização dos dados, como em casos de investigações criminais;
3. Transferência de dados a um outro usuário, respeitando o tratamento de dado;
4. uso dos dados por um controlador, desde que seu acesso seja vedado e os dados sejam anonimizados para uso.

**3.2. 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.

**3.2.1. 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.

Recentemente a Google lançou a versão 2 do TensorFlow, revisando funções e algoritmos como também melhorando implementações, facilitando a implementação dos modelos.

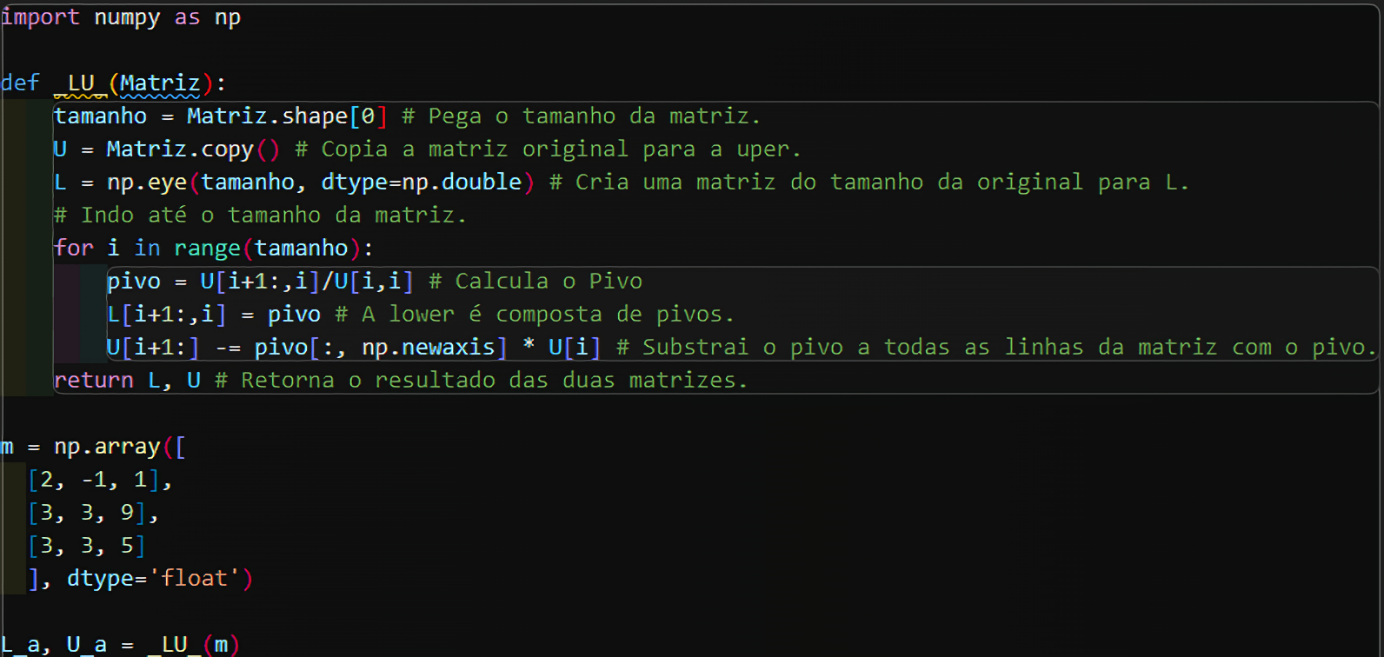
Existem três tipos mais comuns de implementação do TensorFlow, implementado em Python (primeira implementação do framework), JavaScript (Implementação WEB) e TensorFlow Lite, usado em sistemas leves tais quais IOS, Android e Raspberrys Pi’s. em Python, os *tensors* e vértices são os Python Objects e a aplicação TensorFlow é o próprio programa Python. Em JavaScript, utiliza-se a biblioteca WebGL para realizar os processamentos, operações, indexações, sendo o arquivo JS fornecido o TensorFlow e os *tensors* os objetos e funções em WebGL.

Como o WebGL é uma aplicação da Web, fornece a vantagem do TensorFlow pode rodar em qualquer sistema, desde que o navegador do sistema possa interagir com a interface da aplicação Web.

**3.2.2. NUMPY**

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, isso torna NumPy uma biblioteca importante para os cientistas de dados da área utilizando Python, principalmente na área de machine learning do qual os dados estão estruturados para serem utilizados no aprendizado de máquina.

Figura 1 Exemplo de operação com matrizes utilizando NumPy.

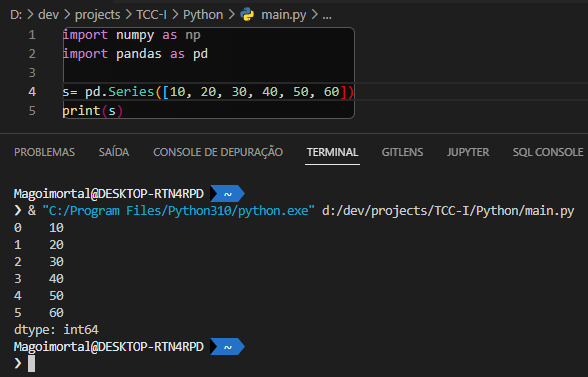


Fonte: Elaborado pelo Autor

**3.2.3. PANDAS**

Desenvolvido como uma API de estrutura de dados, análise de dados e declarado uma biblioteca open-source em 2008, Pandas permite o cientista poder manipular os dados sem perder tempo preparando as estruturas do zero para realização de 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.

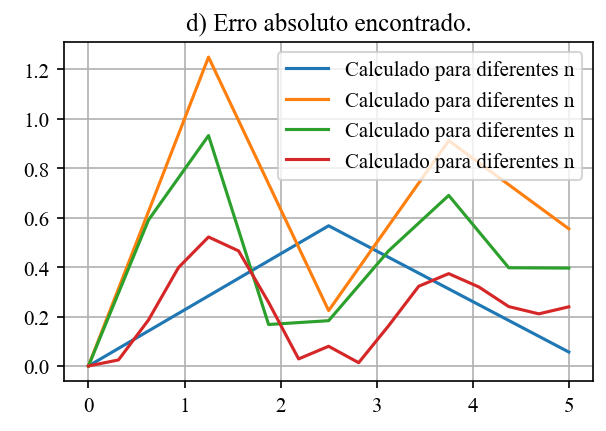


Fonte: Elaborado pelo autor.

**3.2.4. MATPLOTLIB**

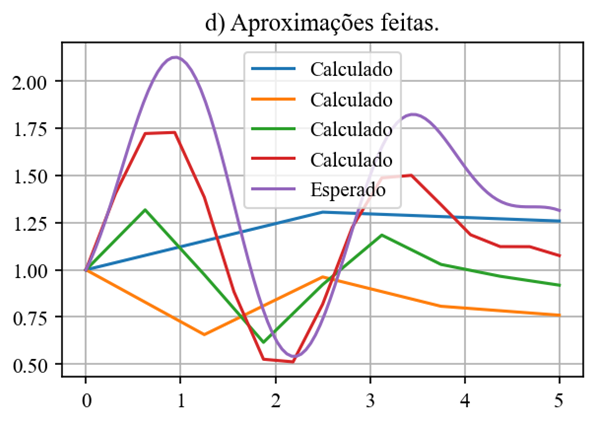
Matplotlib é uma API implementada em Python para visualização de dados, permitindo ver os dados como dados estáticos, animados ou interativos, sendo uma biblioteca poderosa para incorporar junto do NumPy para ver resultados de operações desejadas, permitindo o cientista um poder maior de data visualizada.

Figura 3 [Comparação de erro entre as funções implementadas](https://docs.google.com/document/d/1W2-uly29PnJ2yKRkN4nGQBo6H4uIUrc1DQ9scNfkMyY/edit#heading=h.3dy6vkm)



Fonte: Elaborado pelo Autor

Figura 4 Comparativo entre métodos de Euler criados e o método do NumPy.



Fonte: Elaborado pelo Autor

**3.2.4. MINIMIZAÇÃO DE DADOS**

A minimização de dados será trabalhada em conjunto com o conceito de terminação de dados pessoais pela lei, o modelo do tensorflow será trabalhado e treinado para ser capaz de processar tais dados, com assistência de um terceiro(controlador), sendo então, capaz de deletar dados conforme a necessidade e anonimizar/encriptar outros dados, permitindo apenas o controlador o acesso dentro dos requisitos pela lei.

# SEGURANÇA DA INFORMAÇÃO

**3.2.6. CRIPTOGRAFIA**

A criptografia de dados pessoais na base de dados é importante para garantir, de acordo da L.G.P.D, a anonimização dos dados, dados sensíveis do usuário como senhas, e-mails, endereços e datas de nascimento. As criptografias pesquisadas são criptografias do tipo de “carne moída” e criptográficas simétricas, onde com uma chave gerada é possível encriptar e desencriptar os dados.

No modelo “carne moída” o dado de entrada é processado com rotações e operações de soma com bits, onde o valor de saída “digerido” é uma hash hexadecimal vinculada apenas a aquele valor de entrada, e esse valor hexadecimal pode ser utilizado para realizar as validações sem arriscar vazamento de dados, como exemplo uma API REST trabalhando com validações estaria usando esses valores hexadecimais e armazenando eles na base, as comparações de dados para validação seria hash para hash.

O outro modelo de criptografia simétrica, utilizando uma chave simétrica é possível encriptar os dados e armazená-los na base, desencriptando-se conforme a necessidade do usuário ou de pessoas autorizadas para manipular os dados, como operadores e controladores.

A anonimização pode ocorrer com criptografia dos dados utilizando SHAs ou o RSA para o qual, dados sensíveis não fiquem expostos ou caso ocorra um vazamento de dados pessoais, estes não possam ser facilmente quebrados e expostos à internet.

**3.2.8. HASHING COM SHA-512 PARA ENCRIPTAR DADOS**

A criptografia com hash é um processo de caminho único, onde dada uma entrada qualquer podendo ser uma string, um arquivo ou um dado qualquer, em seguida uma função H(x) realiza o processo de geração de Hash, gerando no resultado um digest (saída) em hexadecimal.

Para a SHA-512 cada saída deve satisfazer três condições de saída no digest sendo elas, distribuição uniforme: indiferente da entrada cada saída produzida deve ser diferente, tamanho fixo de saída: independente da entrada inserido no algoritmo, a saída sempre terá o tamanho de 512 bits em hexadecimal, resistência a colisão: dada duas entradas diferentes, os seus digests resultantes devem variar e não podem serem os mesmos (*"*).

O processo de geração da SHA-512 pode ser divido em 4 partes, formatando a entrada, Inicialização de buffer, Processamento da entrada, realização do digest.

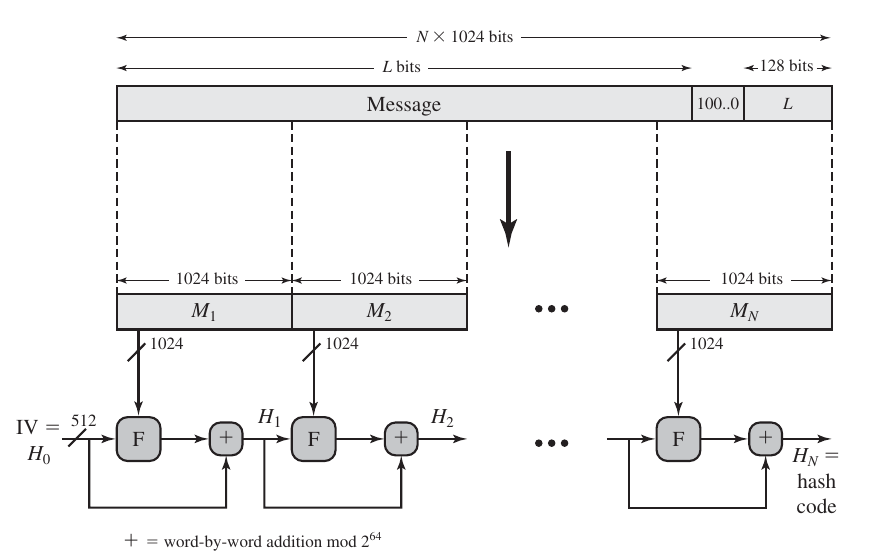
**3.2.8.1 FORMAÇÃO DE ENTRADA**

A entrada é formatada de maneira que o tamanho resultante siga a seguinte equivalência:

Independentemente do tamanho da entrada, mesmo ela já satisfazendo a equivalência da equação acima, a entrada sempre será formatada. Em seguida um bloco de 128 bits é inserido sendo um bloco contém o tamanho original da mensagem, este bloco é tratado como um inteiro de 128 bits de unsigned ints.

O resultado final dessas operações é uma sequência de blocos onde cada bloco possui 1024 bits, de até , resultando em uma mensagem de tamanho N x 1024 bits, como pode ser visualizado na figura abaixo:

Figura 5 Processo de formatação de mensagem



Fonte: Computer Security: Principles and Practices Second Edition.

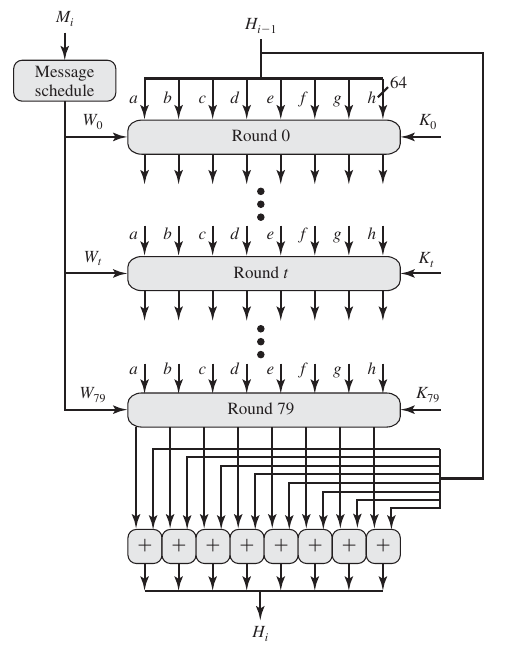
**3.2.8.2 INICIALIZAÇÃO DE BUFFER**

Após a formatação de mensagem, um buffer de 512 bits é iniciado e nele será simulado oito registradores (A, B, C, D, E, F, G, H) cada um com um total de 64 bits de capacidade de armazenamento, cada um dos registradores armazenam os dados em formato de palavra em big-endian (os bits mais significativos estão localizados no lado esquerdo dos bits), essas palavras são obtidas com a raiz quadrada de dos 8 primeiros números primos.

**3.2.8.3 PROCESSAMENTO DE ENTRADA**

O algoritmo procede a realizar um loop em um total de 80, sendo esse loop nomeado de rodadas, em cada rodada o buffer é pego seus registradores e atualiza seus conteúdos. A atualização consiste em utilizar um valor de 64 bits da palavra contida no bloco criado na formatação de entrada, incrementar uma constante aditiva K, está fornecendo um padrão de 64 aleatório para ser utilizado com as operações Booleanas – Depois são feitos rotações, alterações e conjunto de operações Booleanas em bits de AND, OR, NOT e XOR, armazenando o resultado nos registradores, repetindo de 0 até 79 rodadas. No final, pegaremos a o resultado da primeira rodada e adicionaremos a última rodada com um adicional do módulo de .

Figura 6 Processamento de mensagem SHA-512



Fonte: Computer Security: Principles and Practices Second Edition.

**3.2.8.4 REALIZAÇÃO DO DIGEST**

Depois de todo processamento dos blocos esteja completo, obteremos a saída no tamanho definido de 512 bits em hexadecimal.

**3.2.9. ENCRIPTAÇÃO COM RSA PARA DADOS**

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 coprimo do nosso euler totiente e procedemos a calculamos nosso D:

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

Processo de ida:

Processo de volta

# TRABALHOS RELACIONADOS

1. 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>

Nesta implementação da IBM, o QRadar funciona como um sistema automatizado na Cloud para detecção de irregularidades em uma base de dados, caso um cliente deseje seus dados sensíveis deletados, o QRadar vai realizar uma varredura para confirmar se os dados sensíveis foram realmente deletados e ativar um trigger caso este ainda esteja no sistema onde foi detectado. Nesta revisão de literatura, procura entender e ajudar a melhorar o conceito utilizado pela IBM permitindo não só essas detecções, como também avisando quando dados sensíveis não estão encriptados, realizar o descarte de dados sensíveis quando estes não forem descartados direito, invés de apenas avisar.

1. 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>>

Neste trabalho de pesquisa, 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 parametros escolhidos. O projeto a ser criado, propõe não só minimizar os dados como também encriptar os dados sensíveis, sendo uma melhora em relação a esta pesquisa da IBM.

1. **CONCLUSÕES**

A LGPD introduziu na área da Ciências da Computação novos desafios para a área da computação não só pelos requisitos da LGDP, pedindo aos programadores e cientistas para adaptarem o funcionamento de seus softwares e como eles geram dados e quais as necessidades elas precisam para quando forem desenvolvidas. Além disso, a lei gerou um desafio para esses softwares e bases de dados existentes pedindo a adequação deles seguindo a lei.

A área da inteligência artificial, mais especificamente Machine Learning pode ajudar nesses processos, usando treinamento de máquina para entender quais parâmetros inseridos são dados sensíveis e o que fazer com tais dados uma vez detectado, sendo as possíveis ações a minimização, avisos como triggers e anonimização.

Com ferramentas como Python, com o auxílio de NumPy, Pandas e MatPlotLib e JavaScript utilizando WebGL, é possível não só criar e treinar esses modelos utilizando frameworks como o TensorFlow, como também processar e visualizar os resultados, entretanto, Python se demonstrou a ferramenta mais capaz para o desenvolvimento do modelo, não só devido à como ele é utilizado no meio acadêmico/científico como pela sua grande variedade de leques de utilidades para o programador/cientista.

Usando criptografias como Sha-512 e RSA é possível mascarar e anonimizar dados e revertê-los conforme a necessidade do controlador e operador, como apenas eles vão poder ter as chaves privadas para desencriptarem os dados, qualquer terceira parte conseguindo acesso aos dados não vai poder realizar nada com os dados devido a sua encriptação.

Com todos esses conceitos explorados, será desenvolvido um modelo de máquina capaz de analisar bases de dados e possíveis logs para detectar dados sensíveis, descartá-los se forem irrelevantes, e encriptá-los se sensíveis. Permitindo assim uma segurança maior na base e seguir o padrão estabelecido pela lei para dados sensíveis.

# REFERÊNCIA BIBLIOGRÁFICA

**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

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.

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

**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

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

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>

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>

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>

Tankard, Colin; **What the GDPR means for business**, p. 1-8, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1353485816300563?casa\_token=pjOHlq5-iYsAAAAA:dhlGukGxSanwjnBY9aPtw36O1CslOJG1wZ7wTEryTJHi5QtFVM2G6kc8CwuemdVrUDRl2tRqJoA>

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>

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 >

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>

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>

Fatih, Ertam; Galip, Aydin; **Data classification with deep learning using TensorFlow**, p. 1-4, 2017. Disponível em:

<https://ieeexplore.ieee.org/document/8093521>

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 >

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

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

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

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.

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.

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>

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

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.

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>>

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>>

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>

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>>

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>