Universidade Estadual de Londrina

FELIPE ALVES BARUSSO

APLICANDO COMPARTILHAMENTO DE SEGREDOS PARA TREINAR MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM DADOS MÉDICOS COM PRESERVAÇÃO DE PRIVACIDADE

LONDRINA

2023

FELIPE ALVES BARUSSO

APLICANDO COMPARTILHAMENTO DE SEGREDOS PARA TREINAR MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM DADOS MÉDICOS COM PRESERVAÇÃO DE PRIVACIDADE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Bruno Bogaz Zarpelão

LONDRINA

2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Barusso, Felipe Alves.
Aplicando compartilhamento de segredos para treinar modelos de aprendizado de máquina baseados em dados médicos com preservação de privacidade / Felipe Alves Barusso. - Londrina, 2023.
43 f.

Orientador: Bruno Bogaz Zarpelão.
Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Graduação em Ciência da Computação, 2023.
Inclui bibliografia.

1. Aprendizado de máquina - TCC. 2. Privacidade - TCC. 3. Criptografia - TCC. 4. Compartilhamento de segredo - TCC. I. Zarpelão, Bruno Bogaz. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Graduação em Ciência da Computação. III. Título.

CDU 519

FELIPE ALVES BARUSSO

APLICANDO COMPARTILHAMENTO DE SEGREDOS PARA TREINAR MODELOS DE APRENDIZADO DE MÁQUINA BASEADOS EM DADOS MÉDICOS COM PRESERVAÇÃO DE PRIVACIDADE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Orientador: Prof. Dr. Bruno Bogaz Zarpelão
Universidade Estadual de Londrina

Prof. Dr. Gilberto Fernandes Junior
Universidade Estadual de Londrina – UEL

Vitor Castro
Universidade Estadual de Londrina – UEL

Londrina, 3 de novembro de 2023.

AGRADECIMENTOS

Sou grato principalmente ao meu orientador Bruno, por sua persistência e paciência comigo. Agradeço também à minha família que me apoiou e me deu a oportunidade e base para conquistar tudo que tenho hoje. Por fim, agradeço aos amigos, por tornarem a jornada agradável, especialmente Claudio, Gabriel, Guilherme, Luis Felipe, Melvi, Vinícius Cesar e Vinícius Luciano.

"Pois que se uniu a mim, eu o livrarei; e o protegerei, pois conhece o meu nome."
(Bíblia Sagrada, Salmos 91, 14)

BARUSSO, F. A.. **Aplicando compartilhamento de segredos para treinar modelos de aprendizado de máquina baseados em dados médicos com preservação de privacidade**. 2023. 43f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

Nos serviços médicos atuais, os dados de um indivíduo são muitas vezes armazenados digitalmente. Com estas informações, é possível aplicar algoritmos de aprendizado de máquina para automatizar processos e realizar análises sobre elas. Como estas operações podem demandar recursos computacionais mais especializados e robustos, é possível que instituições de saúde aloquem estas tarefas para servidores externos que atendam estes requisitos. Contudo, devido à natureza sensível e pessoal das informações contidas nestas bases de dados, é necessário serem implementadas medidas de garantia de privacidade para que elas possam ser repassadas para estes servidores externos. Neste trabalho, foi considerado o cenário de uma instituição de saúde realizar o treinamento de um modelo de classificação utilizando servidores remotos, porém protegendo seus dados de treinamento. Este cenário foi então simulado com o treinamento dividido em múltiplos servidores de um modelo de regressão logística sobre dados médicos utilizando preservação de privacidade. Foi treinado também um modelo de regressão logística de maneira convencional, e ambos foram analisados para avaliar o impacto da preservação de privacidade na capacidade preditiva do modelo. A simulação foi repetida múltiplas vezes e utilizando dois conjuntos de dados diferentes para evitar enviesamento dos resultados. Os resultados mostraram não haver impacto significativo na qualidade de um modelo treinado com o método de preservação de privacidade escolhido.

Palavras-chave: Aprendizado de máquina. Privacidade. Criptografia. Compartilhamento de segredo.

BARUSSO, F. A.. **Applying secret sharing to train machine learning models based on medical data with privacy preservation**. 2023. 43p. Final Project (Bachelor of Science in Computer Science) – State University of Londrina, Londrina, 2023.

ABSTRACT

In today's medical services, an individual's data is often stored digitally. With this information, it is possible to apply machine learning algorithms to automate processes and perform analysis on them. As these operations may demand more specialized and robust computational resources, it is possible for health institutions to offload these tasks to external servers that meet these requirements. However, due to the sensitive and personal nature of the information contained in these databases, it is necessary to implement privacy guarantee measures so that they can be passed on to these external servers. In this work, we considered the scenario of a health institution training a classification model using remote servers, but protecting its training data. This scenario was then simulated by training a logistic regression model on medical data using privacy preservation on a multiserver environment. A logistic regression model was also trained conventionally, and both were analyzed to assess the impact of privacy preservation on the model's predictive capabilities. The simulation was repeated multiple times and using two different datasets to avoid possible bias. The results showed no significant impact on the quality of a model trained with the chosen privacy preservation method.

Keywords: Machine learning. Privacy. Cryptography. Secret sharing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LGPD	*Lei Geral de Proteção de Dados*
GDPR	*General Data Protection Regulation*
MPC	*Secure Multi-Party Computation*
TFE	*TF Encrypted*
BCW	*Breast Cancer Wisconsin*
PID	*Pima Indians Diabetes*
MCC	*Matthew's Correlation Coefficient*

SUMÁRIO

1 INTRODUÇÃO

A transição dos serviços de saúde para o meio digital apresentou benefícios como processos mais eficientes e a diminuição de custos, levando ao surgimento de novos serviços e modelos de negócios [1]. Com estas novas aplicações, informações sobre consultas, dados de saúde e imagens de exames de um paciente são armazenadas nos bancos de dados de instituições médicas.

Com estes dados, é possível aplicar técnicas de aprendizado de máquina para apoiar atividades médicas. Uma possível aplicação é um algoritmo que automatiza parte de um diagnóstico. Outro exemplo, é um modelo de classificação que identifica tecidos malignos em imagens de núcleos de células mamárias [2].

Porém, são grandes volumes de informação sendo processados por algoritmos que normalmente demandam servidores especializados. As instituições podem optar por criar arquiteturas de sistema onde o processamento destas tarefas mais custosas seja realizado em servidores externos na nuvem, por exemplo. O problema é que espalhar dados sensíveis em diferentes servidores e diferentes ambientes aumenta o risco de que haja vazamentos. Com isso, é necessário implementar medidas que garantam a privacidade dos dados.

Dados médicos frequentemente contém informações que podem revelar a identidade de um paciente, tornando o controle de acesso aos mesmos um fator importante. Além disso, os dados de saúde pessoal têm um alto valor para as empresas farmacêuticas, seguradoras e empregadores, o que os tornam alvo de ataques [3].

Nem mesmo empresas com um alto investimento em segurança estão imunes a vazamento de dados. Em agosto de 2021, a empresa americana de segurança em nuvem Wiz encontrou uma vulnerabilidade crítica na plataforma Microsoft Azure, que permitia o controle total e remoto de contas de outros usuários do banco de dados da Azure, o Cosmos DB [4]. Alguns clientes do Cosmos DB incluem: Coca-Cola, ExxonMobil e Walgreens.

Uma possível solução para a preservar a privacidade da informação ao enviá-la para servidores externos é o conceito de compartilhamento de segredo [5]. A instituição que deseja fazer o uso de ambientes externos de processamento primeiro divide os dados e acrescenta um ruído a cada parte resultante. Então, a instituição proprietária dos dados envia cada uma dessas partes a um servidor externo, que não têm contato com o conjunto de dados original. Por fim, os servidores externos realizam as operações requisitadas sobre os dados modificados pelo ruído (dados codificados). Os resultados codificados são enviados de volta para o dono dos dados, que deve reuni-los e decodificá-los para revelar o resultado desejado.

Neste trabalho, foi utilizada uma técnica de compartilhamento de segredo para rea-

lizar o treinamento de modelos de classificação com regressão logística utilizando múltiplos servidores terceiros. Os modelos foram treinados em dados médicos, os quais continham informações sensíveis de exames de pacientes. Para cada modelo treinado com preservação de privacidade, um segundo modelo de regressão logística foi treinado com o mesmo conjunto de dados, mas sem preservação de privacidade. Assim, buscou-se avaliar se a implementação da preservação de privacidade no treinamento é viável com base em seu impacto na qualidade preditiva dos modelos. Os resultados obtidos não demonstraram degradação notável na predição dos modelos que realizaram o treinamento utilizando o compartilhamento de segredo.

O restante deste trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os conceitos de aprendizado de máquina, suas categorias e as técnicas que serão relevantes para os experimentos, bem como os desafios da privacidade na área de tecnologia aplicada a saúde. O Capítulo 3 mostra a modelagem do problema proposto, os materiais e os métodos que serão utilizados. O Capítulo 4 discorre sobre os experimentos em si, os resultados, e as capacidades dos modelos treinados. Por fim, o Capítulo 5 fala sobre as implicações dos resultados obtidos no contexto geral do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aprendizado de máquina

2.1.1 Visão geral

O aprendizado de máquina é a ciência de programar computadores para eles poderem aprender com dados [6]. O aprendizado ocorre por algoritmos e modelos estatísticos que realizam uma tarefa específica sem serem explicitamente programados para tal fim [7]. Em outras palavras, o aprendizado de máquina permite que os computadores identifiquem padrões automaticamente e realizem previsões com base em dados de entrada, sem intervenção humana. Isso é feito por meio do uso de conjuntos de dados e algoritmos projetados para identificar relacionamentos e padrões nesses dados.

Existem vários tipos de aprendizado de máquina, incluindo aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e aprendizado por reforço. Cada uma dessas abordagens tem objetivos e técnicas diferentes, mas todas compartilham o objetivo comum de permitir que as máquinas aprendam com um determinado conjunto de dados [8].

O aprendizado de máquina tem inúmeras aplicações em uma ampla gama de setores, incluindo saúde, finanças, transporte e entretenimento. Alguns exemplos de aprendizado de máquina em ação incluem reconhecimento de imagem, processamento de linguagem natural, sistemas de recomendação e análise preditiva [9].

2.1.2 Aprendizado supervisionado

O aprendizado supervisionado é responsável por uma parte expressiva da pesquisa na área de aprendizado de máquina. O nome sugere a ideia de um 'supervisor' que instrui o modelo de aprendizado sobre os rótulos a serem associados com exemplos de treinamento. Algoritmos de aprendizado supervisionado induzem modelos a partir desses dados de treinamento e esses modelos podem ser usados para classificar outros dados não rotulados [10].

Uma base de dados empregada em uma tarefa de aprendizagem é composta por um grupo de registros descritos por um conjunto de atributos $A = \{a_1, a_2, ..., a_{|A|}\}$, onde $|A|$ denota o número de atributos do conjunto A. Essa base de dados também possui um rótulo C, chamado de atributo de classe. É assumido que C não pertence a A e tem um conjunto de valores $C = \{c_1, c_2, ..., c_{|C|}\}$, onde $|C|$ é o número de classes e $|C| >= 2$ [11].

Um conjunto de dados para aprendizado é simplesmente uma tabela relacional, como demonstra a Tabela 1. Dado um conjunto de dados D, o objetivo do aprendizado é

produzir uma função de classificação para relacionar valores de atributos em A e classes em C. A função pode ser usada para prever os rótulos de classe de dados futuros, sendo também chamada de modelo de classificação.

Tabela 1 – Exemplo do conjunto de dados Iris, que contém informações sobre plantas da família *Iridaceae*, cada instância classificada em três possíveis gêneros. (Fonte: 12)

sepal_lenght	sepal_width	petal_lenght	petal_width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.0	3.0	1.4	0.2	Iris-setosa
. 7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.8	1.9	Iris-virginica

Depois que um modelo é treinado a partir dos dados de treinamento por um algoritmo de aprendizado, ele pode ser avaliado usando um conjunto de dados de teste para medir a precisão do modelo. É importante observar que os dados de teste não são usados no aprendizado do modelo de classificação. Os exemplos nos dados de teste também possuem rótulos de classe, e é por isso que eles podem ser usados para avaliar o modelo treinado, por ser possível verificar se a classe prevista para cada caso de teste pelo modelo é a mesma que a classe real do caso de teste [13].

2.1.3 Aprendizado não supervisionado

O aprendizado não supervisionado é uma técnica de aprendizado de máquina que permite que o modelo aprenda a partir de dados que não possuem rótulos ou categorias pré-definidas. Nessa abordagem, o modelo consegue identificar padrões e estruturas por conta própria, sem a necessidade de informações externas [14].

Ao contrário do aprendizado supervisionado, no qual o modelo é treinado com exemplos rotulados para aprender a relação entre os recursos de entrada e a saída desejada, o aprendizado não supervisionado é utilizado para descobrir a estrutura latente nos dados. Essa técnica é especialmente útil para tarefas como *clustering* (agrupamento) de dados não rotulados, redução de dimensionalidade e descoberta de padrões em dados [15].

Entre as técnicas mais comuns de aprendizado não supervisionado, é possível citar o *clustering*, a análise de componentes principais, a análise de fatores e a análise de conglomerados [16].

2.1.4 Aprendizado semi-supervisionado

O aprendizado semi-supervisionado consiste no treinamento de um modelo com uma combinação de dados rotulados e não rotulados. A ideia é que o modelo use os

dados rotulados para aprender com exemplos explícitos e generalizar para novos pontos de dados, enquanto os dados não rotulados ajudam o modelo a descobrir e capturar a estrutura subjacente dos dados [17].

Uma abordagem comum para o aprendizado semi-supervisionado é usar uma pequena quantidade de dados rotulados para treinar um modelo e, em seguida, usar o modelo para fazer previsões sobre os dados não rotulados. Os rótulos previstos são então usados para atualizar o modelo, treinado novamente nos dados rotulados e recém-rotulados. Este processo pode ser repetido até que o modelo atinja um nível satisfatório de precisão.

2.1.5 Aprendizado por reforço

O aprendizado por reforço é baseado no conceito de aprendizado por tentativa e erro. Seu objetivo é permitir que um agente aprenda como tomar decisões que maximizem um sinal de recompensa cumulativo. No aprendizado por reforço, o agente interage com um ambiente, o qual fornece um *feedback* na forma de recompensas ou punições com base nas ações tomadas [18].

A cada intervalo de tempo, o agente observa o estado atual do ambiente e seleciona uma ação com base em seu conhecimento atual. O ambiente então faz a transição para um novo estado e o agente recebe um sinal de recompensa com base no novo estado e na ação realizada. O agente então atualiza seu conhecimento com base no sinal de recompensa e no novo estado [19].

2.1.6 Regressão linear

A regressão linear é um método de aprendizado supervisionado. Ela consiste em um modelo estatístico que assume uma relação linear entre um conjunto de variáveis dependentes, que se visa prever, e um conjunto de variáveis independentes, que se visa avaliar. A regressão linear produz uma equação matemática (ou modelo) para uma linha de melhor ajuste para descrever a relação [20].

Em outras palavras, modelos lineares podem ser usados para modelar a dependência de uma variável dependente y em relação a um conjunto de variáveis independentes x. No caso do aprendizado supervisionado, x representa os atributos de classe e y os rótulos. As relações aprendidas são lineares e podem ser escritas para uma única instância i da seguinte forma [21]:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_p\beta_p + \epsilon \tag{2.1}$$

O resultado previsto de uma instância é uma soma ponderada de suas p características. $\beta_0 ... \beta_p$ representam os pesos ou coeficientes de características aprendidas. O primeiro peso na soma (β_0) é chamado de intercepto e não é multiplicado por um recurso.

O épsilon (ϵ) é o erro cometido, ou seja, a diferença entre a previsão e o resultado real [21].

Vários métodos podem ser usados para estimar os pesos ideais. O método dos mínimos quadrados ordinários é geralmente usado para encontrar os pesos que minimizam as diferenças quadradas entre os resultados reais e estimados [21]:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2 \tag{2.2}$$

Onde $\hat{\beta}$ é o conjunto de pesos ajustados, n é o número de elementos utilizados para treinamento, $y^{(i)}$ é a variável dependente de índice i (relacionada ao conjunto de variáveis independentes $x^{(i)}$) e $\beta_0 ... \beta_p$ é o conjunto de pesos.

2.1.7 Regressão logística

A regressão logística modela as probabilidades para problemas de classificação com dois resultados possíveis. É uma extensão do modelo de regressão linear para problemas de classificação. Em vez de ajustar uma linha reta ou hiperplano, o modelo de regressão logística usa a função logística para encaixar a saída de uma equação linear entre 0 e 1. A função logística é definida como [21]:

$$\mathrm{l}(\eta) = \frac{1}{1 + \exp(-\eta)} \tag{2.3}$$

Este método tem origem na regressão linear, então partimos da equação linear 2.1 (com exceção do épsilon). Porém, para tarefas de classificação, são preferíveis probabilidades entre 0 e 1. Então o lado direito da Equação 2.1 é envolvido com a função logística. Isso força a saída a assumir apenas valores entre 0 e 1 [21]:

$$P\left(y^{(i)} = 1\right) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}\right)\right)} \tag{2.4}$$

Por fim, uma técnica para obter o conjunto de pesos w (composto por $\beta_0 ... \beta_p$) que melhor descrevem a relação logística entre o conjunto de entrada X e os rótulos y é minimizar uma função de custo, como, por exemplo, a de entropia cruzada, descrita na Equação 2.5:

$$C(w) = \frac{1}{N} \sum_{i=1}^{N} \left(-y_i \log s(X_i \cdot w) - (1 - y_i) \log \left(1 - s(X_i \cdot w)\right) \right) \tag{2.5}$$

Onde N é o número de entradas, X_i é o elemento i do conjunto X, y_i é o rótulo deste elemento, w é o vetor de pesos e s é a função sigmoide $s(z) = 1/(1 + e^{-z})$

Para minimizar o valor da função de entropia cruzada e obter um vetor de pesos w mais preciso é possível utilizar o método do gradiente, recursivamente atualizando o vetor de pesos em cada iteração do treinamento, como demonstra a Equação 2.6.

$$w^{(t+1)} = w^{(t)} - \frac{\alpha}{N} X^{\top} \left(s\left(X \times w^{(t)} \right) - y \right) \tag{2.6}$$

Onde $w^{(t)}$ é o vetor de pesos na iteração t da etapa de treinamento, α é a taxa de aprendizagem e $X^{\top}$ é a matriz transposta do conjunto de entrada.

2.2 Privacidade

As preocupações com a privacidade em relação à coleta, armazenamento e uso de informações pessoais são tópicos recorrentes na sociedade atual. Atualmente, a análise de dados está profundamente incorporada em nossos sistemas. Essas análises, individualmente ou em combinação com outras, podem revelar atributos sensíveis dos indivíduos, incluindo informações sobre sua saúde, finanças, tendências políticas e comportamentos sociais [22].

A Lei Geral de Proteção de Dados (LGPD) [23] é uma lei brasileira que entrou em vigor em setembro de 2020 e estabelece as regras para que as companhias e instituições possam coletar, guardar, usar e compartilhar informações pessoais dos indivíduos. A LGPD protege a privacidade e os direitos dos titulares dos dados pessoais, garantindo que as informações sejam usadas de forma transparente, legítima e segura. Ela estabelece regras claras para a coleta, o uso e o compartilhamento de informações pessoais, além de estabelecer punições em caso de violação.

Diante dos riscos aos indivíduos, as organizações que gerenciam dados pessoais devem implementar várias medidas para proteger a privacidade daqueles cujas informações pessoais são usadas. No entanto, o cenário de privacidade de dados está se modificando depressa, à medida que os avanços tecnológicos alteram como as informações pessoais são coletadas, armazenadas e compartilhadas [24].

As expectativas sobre privacidade estão mudando significativamente, em resposta a um amplo espectro de políticas invasivas nocivas à privacidade subjacentes a aplicativos no uso diário. Devido a essas políticas, medidas técnicas são implantadas para mitigar os riscos de privacidade informacional. Essas medidas incluem adição de ruído, dados sintéticos e criptografia [25].

Com essa dependência de abordagens tecnológicas, a privacidade vem se tornando cada vez mais um conceito técnico. Além disso, a compreensão da privacidade refletida nessas tecnologias está evoluindo, em parte em resposta à descoberta de novas vulnerabilidades. Os principais exemplos que ilustram a evolução das práticas de privacidade podem

ser encontrados nas políticas de agências federais, que buscaram fortalecer as proteções em resposta às ameaças à privacidade.

2.2.1 Privacidade na área de tecnologia aplicada à saúde

A aquisição, armazenamento e uso de informações pessoais de saúde são processos necessários para muitas atividades básicas de saúde pública [26]. As preocupações com a confidencialidade geraram discussões sobre o equilíbrio ideal entre os interesses individuais e sociais.

Preocupações relativas à confidencialidade tornaram-se ainda maiores na era digital. Violações de alto perfil das informações de saúde dos indivíduos aumentaram a ansiedade sobre a privacidade. Além disso, planos para criar redes interconectadas de informações de saúde têm gerado discussões acaloradas em torno da privacidade dos dados [27].

Na esfera da saúde pública, várias violações amplamente divulgadas ocorreram nos últimos anos. Elas incluem o vazamento de informações pessoais de cerca de 5 milhões de militares americanos após o roubo de fitas de backup de registros eletrônicos do programa de saúde americano *Tricare* [28], o ataque ao grupo *Shields Healthcare Group* que comprometeu dados pessoais de cerca de 2 milhões de pacientes em 2022 [29], entre outros.

Uma das razões pelas quais a Comissão Europeia desenvolveu o Regulamento Geral de Proteção de Dados (GDPR), que se tornou uma diretiva regulatória para a União Europeia, foi o obstáculo da privacidade para a realização da promessa da saúde digital. O GDPR reconhece que os indivíduos precisam controlar seus próprios dados, mas também declara a necessidade de confiança nos serviços de dados pessoais por meio transparência e o emprego de tecnologias seguras [1].

Conforme o GDPR, as organizações devem implementar medidas tecnológicas e operacionais adequadas para proteger os dados, incluindo a adoção de fortes controles de privacidade. Ele também afirma que as organizações devem adotar medidas internas que atendam aos princípios de proteção de dados por design. Na prática, isso significa que a proteção de dados e privacidade devem ser consideradas desde o início do processo de planejamento de segurança [30].

2.2.2 Ameaças à privacidade no aprendizado de máquina

O impacto do aprendizado de máquina na segurança, privacidade e imparcialidade está recebendo cada vez mais atenção. Dados pessoais são coletados por diversos serviços online, sendo utilizados para treinar modelos de aprendizado de máquina. No entanto, não se sabe se tais modelos revelam informações sobre os dados utilizados para

seu treinamento. Se um modelo é treinado utilizando dados confidenciais, como localização, registros de saúde ou informações de identidade, um ataque que permita que um adversário extraia essas informações do modelo é altamente indesejável [31].

Segundo Biggio e Roli [32], os ataques a sistemas de aprendizado de máquina são divididos em três categorias: ataques contra a integridade, como ataques de evasão, que causam erros de classificação de amostras específicas; ataques de disponibilidade, como técnicas de envenenamento, que buscam maximizar o erro de classificação e ataques contra privacidade e confidencialidade, ou seja, ataques que tentam inferir informações sobre dados de usuários ou sobre a estrutura de um modelo.

Para compreender e se proteger contra ataques sob a perspetiva da privacidade, é útil ter um modelo do ambiente, dos diferentes atores e dos recursos que devem ser protegidos. Do ponto de vista do modelo de ameaça, os recursos sensíveis e potencialmente sob ataque são o conjunto de dados de treinamento, o próprio modelo, seus parâmetros e sua arquitetura. Os atores são:

1. Os **proprietários de dados**, cujos dados podem ser confidenciais.
2. O **proprietário do modelo**, que pode ou não querer compartilhar informações sobre o modelo.
3. Os **consumidores do modelo**, que utilizam os serviços que o proprietário do modelo expõe.
4. Os **adversários**, que também podem ter acesso às interfaces do modelo como um consumidor normal. Se o proprietário do modelo permitir, eles poderão ter acesso ao próprio modelo.

A Figura 1 mostra o ambiente e os atores relevantes de um ataque a um sistema que utiliza aprendizado de máquina. É importante ressaltar que ela apresenta um modelo lógico e não impede a possibilidade de que alguns desses atores ou recursos possam estar separados em múltiplos locais, ou de um ator exercer mais de um papel.

As diferentes técnicas de ataque contra modelos de aprendizado de máquina podem ser modeladas em termos de conhecimento adversário. Ataques onde o adversário não tem conhecimento dos parâmetros do modelo, arquitetura ou dados de treinamento são chamados de ataques de caixa preta. Por outro lado, ataques nos quais o adversário tem conhecimento de tais recursos são conhecidos como de ataques caixa branca [33].

Além disso, o comportamento de um adversário também é considerado na modelagem. Um adversário semi-honesto (passivo), não interfere no treinamento e tenta apenas inferir conhecimento durante ou após o treinamento. Se o adversário interfere de alguma forma no processo de treinamento, ele é considerado um adversário malicioso (ativo) [33].

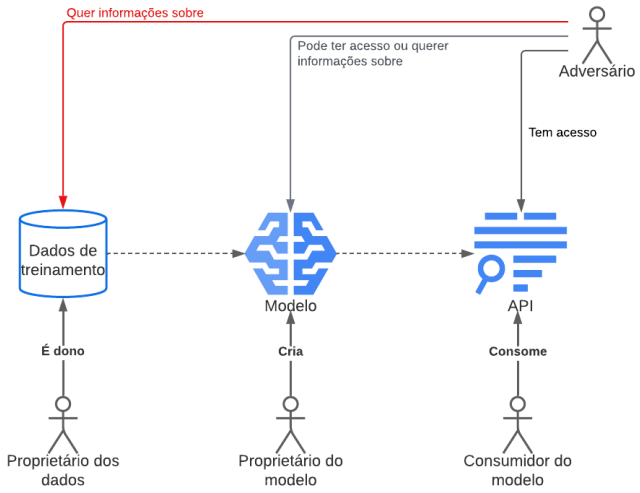

Figura 1 – Modelo do ambiente e dos atores de um ataque à uma aplicação de aprendizado de máquina.

2.2.2.1 Ataques de inferência de associação

Um ataque de inferência de associação tenta determinar se uma amostra de entrada foi usada como parte do conjunto de treinamento. Uma das condições que demonstra melhorar a precisão desse tipo de ataque é a má generalização de um modelo. Isso ocorre quando há um sobre-ajuste no modelo, ou seja, no treinamento ele se ajusta muito bem ao conjunto treinado, mas é ineficaz com entradas novas. Outro fator que aumenta a precisão de um ataque de inferência de associação é o acesso aos parâmetros e gradientes do modelo durante o treinamento, ou seja, ataques caixa branca em geral são mais eficientes [34].

A defesa mais proeminente contra ataques de inferência de associação é a Privacidade Diferencial, que fornece garantias sobre o impacto que os registros de dados individuais tenham na saída de um modelo. Este método consiste na adição de ruído (Laplaciano ou Gaussiano, por exemplo) à saída de uma consulta, ou função no conjunto de dados. A quantidade de ruído adicionado a uma função é determinada pelo parâmetro de privacidade, chamado ϵ (épsilon). Quanto menor o valor de ϵ, maior a privacidade dos dados, mas maior será o ruído adicionado à função. O parâmetro ϵ controla a quantidade de aleatoriedade adicionada à função para garantir que os resultados não revelem informações sensíveis sobre os dados de entrada [35]. Além disso, é possível se proteger de ataques de inferência evitando o sobre-ajuste do modelo, através, por exemplo, do uso de conjuntos de validação.

2.2.2.2 Ataques de reconstrução

Os ataques de reconstrução tentam recriar uma ou mais amostras de treinamento e/ou seus respectivos rótulos de treinamento. A reconstrução pode ser parcial ou total.

Fatores que tornam um modelo mais suscetível a um ataque de reconstrução incluem a alta correlação entre os dados de entrada e uma alta sensibilidade (pequenas mudanças nos dados de entrada podem ter um grande impacto nos resultados) de um modelo [36].

Os ataques de reconstrução geralmente requerem acesso aos gradientes calculados na função de perda durante o treinamento. A maioria das defesas contra ataques de reconstrução propõe técnicas que afetam as informações recuperadas desses gradientes. Um método comum é definir todos os gradientes de perda abaixo de um certo limite como zero [37].

2.2.2.3 Ataques de inferência de propriedade

A capacidade de obter informações sobre o conjunto de dados que não estão diretamente correlacionadas com a tarefa de aprendizado é chamada de inferência de propriedade. Um exemplo é a extração de informações sobre a proporção de mulheres e homens em um conjunto de dados de pacientes quando essa informação não era um atributo codificado ou um rótulo do conjunto de dados. A privacidade diferencial também é um método de defesa contra este tipo de ataque [38].

2.2.3 Computação multipartidária segura

A computação multipartidária segura (ou MPC, do inglês *Secure Multi-Party Computation*) define protocolos que permitem que múltiplas partes colaborem na realização de um cálculo ou processamento de dados, sem divulgar informações confidenciais umas às outras. Seu objetivo é garantir a privacidade e segurança dos dados, mesmo quando várias partes precisam compartilhar informações sensíveis para realizar uma tarefa específica [39].

A MPC tem aplicações em diversos contextos, incluindo eleições, saúde, bancos e comércio eletrônico. Por exemplo, a MPC pode ser usada em eleições para garantir que os votos sejam contados de forma justa e precisa, sem expor a escolha dos eleitores a ninguém. Da mesma forma, a MPC pode ser usada em ambientes bancários para permitir que vários bancos realizem cálculos de risco colaborativamente, sem expor informações sensíveis a outros bancos.

Embora a MPC seja considerada altamente segura, sua eficiência pode variar dependendo do tamanho do problema a ser resolvido e do número de participantes envolvidos [40].

2.2.4 Compartilhamento de segredo

Em um protocolo MPC, as partes precisam compartilhar suas entradas privadas entre si sem revelá-las. Uma maneira de atingir este objetivo é o compartilhamento de

segredo. O compartilhamento de segredo é uma técnica criptográfica que permite que uma informação seja dividida em partes e distribuída para diferentes pessoas ou entidades, garantindo que somente a combinação das partes permita a recuperação da informação original [41].

O compartilhamento de segredo é implementado por meio de um algoritmo que divide a informação em "segredos compartilhados", sendo distribuídos entre as partes envolvidas. Esses segredos são projetados de forma que a informação original só possa ser recuperada se um número mínimo de segredos for combinado. Em um esquema de compartilhamento de segredos, uma informação é dividida em n partes (uma para cada participante) e o número de partes que são necessárias para reconstruir o segredo é chamado de k.

Existem diferentes algoritmos para implementar o compartilhamento de segredo, incluindo o esquema de Shamir [5], que consiste em ajustar um polinômio de grau $k-1$ a qualquer conjunto de k pontos que estejam no polinômio, sendo necessários t pontos para definir um polinômio de grau $k-1$. O objetivo é criar um polinômio de grau $k-1$ com o segredo como o primeiro coeficiente e os demais coeficientes escolhidos aleatoriamente. Em seguida, é necessário escolher n pontos na curva e enviá-los para cada participante. Quando pelo menos k dos n jogadores revelam seus pontos, há informação suficiente para ajustar um polinômio de grau $k-1$ a eles, sendo o primeiro coeficiente o segredo.

Outra técnica notável é o esquema de Blakley [42]. Ela parte do princípio de que todo conjunto de n hiperplanos de dimensão $n-1$ se interceptam em um ponto único. O segredo pode ser codificado como qualquer coordenada única do ponto de interseção. Cada participante recebe informações suficientes para definir um hiperplano; o segredo é recuperado calculando o ponto de interseção dos planos e, em seguida, tomando uma coordenada especificada dessa interseção.

É importante destacar que, embora o compartilhamento de segredos possa ser uma técnica poderosa para proteger informações confidenciais, sua segurança depende da escolha adequada do algoritmo de compartilhamento e do número mínimo de partes necessárias para recuperar a informação original.

Conforme mostra a Figura 2, um esquema de compartilhamento de segredo em n partes consiste em dois processos: distribuição e reconstrução. O processo de distribuição recebe como entrada o segredo S e gera n partes $S_1, S_2, S_3, ..., S_N$ que são entregues privadamente (utilizando, por exemplo, adição de ruído) aos participantes do sistema. O processo de reconstrução, normalmente feito por uma parte mestra, reconstrói o segredo correto quando provido com qualquer subconjunto de k compartilhamentos. Um algoritmo de compartilhamento de um segredo S é denotado por $A(S)$ e as n partes resultantes por $S_1, S_2, S_3, ..., S_N$ [43].

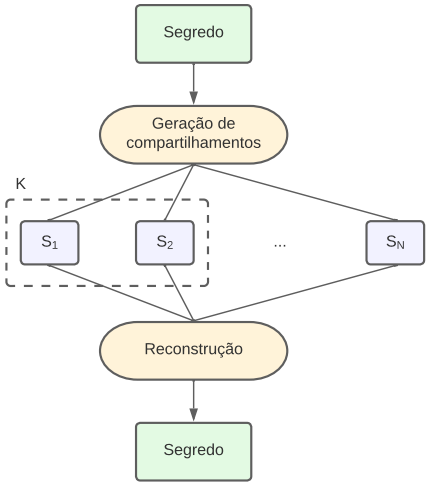

Figura 2 – Esquema de compartilhamento de segredos.

2.2.5 Protocolo ABY³

O ABY³ é um *framework* implementado em C++ para aprendizado de máquina que preserva a privacidade e apresenta soluções para treinar regressão linear, regressão logística e modelos de rede neural. Ele define protocolos em um modelo de três servidores onde os proprietários de dados compartilham seus dados entre os servidores através do compartilhamento de segredo, treinando e avaliando modelos nos dados conjuntos usando MPC. O *framework* executa cálculos alternando entre diferentes tipos de técnicas de compartilhamento de segredos [44].

No ABY³, cada entrada privada $x \in Z_{2^k}$ (onde k é o número de bits utilizado para representar um número, por exemplo, $k = 64$) é secretamente compartilhado com três partes P_0, P_1, P_2, usando uma das três seguintes técnicas (embora na regressão logística, pertinente a este trabalho, o protocolo utilize apenas a primeira):

1. ***A**rithmetic secret sharing*: Selecione três valores aleatórios $x_0, x_1, x_2 \in Z_{2^k}$, tais que $x = x_0 + x_1 + x_2$. Cada parte P_i detém x_i e $x_{i+1 mod 3}$. Este compartilhamento é denotado como $[[x]]^A$.

2. ***B**inary secret sharing*: Selecione três valores aleatórios $x_0, x_1, x_2 \in Z_{2^k}$, tais que $x = x_0 \bigoplus x_1 \bigoplus x_2$. Cada parte P_i detém x_i e $x_{i+1 mod 3}$. Este compartilhamento é denotado como $[[x]]^B$.

3. ***Y**ao sharing*: Aplica o protocolo *Yao's garbled circuits* [45] com P_0 atuando como avaliador e P_1 e P_2 atuando como *garblers*. Este compartilhamento é denotado como $[[x]]^Y$.

3 MATERIAIS E MÉTODOS

3.1 Visão geral

Neste trabalho é proposto um cenário hipotético onde um hospital decide realizar o treinamento de um modelo de classificação utilizando dados médicos sensíveis. Para isso, foram definidas duas condições:

1. **O treinamento deve ser realizado remotamente**. De forma geral, para obter um modelo de classificação melhor, é necessário acrescer a massa de dados de entrada. Conforme o tamanho do conjunto de dados de treinamento cresce, o custo computacional da etapa de treinamento aumenta. Assim, faz sentido que a instituição busque realizar este processo remotamente, em servidores projetados para realizarem operações mais custosas.
2. **É necessário empregar técnicas de criptografia no processo de treinamento**. Como o conjunto de dados de treino contém informações sensíveis pessoais, a instituição procura se proteger contra ataques ou um eventual vazamento de dados.

Para simular este cenário, foi escolhida a biblioteca *python* TF Encrypted (TFE) [46]. A TFE é um *framework* de código aberto, construída com base na conhecida biblioteca *TensorFlow*, para cálculos com preservação de privacidade. Ela possui uma interface que segue os padrões da *TensorFlow*, permitindo que usuários que não são especialistas em criptografia desenvolvam soluções de aprendizado de máquina seguras. A biblioteca suporta diversos protocolos de criptografia e MPC, mas neste trabalho foi utilizado apenas o protocolo ABY^3.

Assim, conforme ilustrado na Figura 3, o cenário é modelado da seguinte forma: o hospital é o detentor de um conjunto X de dados de treinamento. Utilizando as funcionalidades da TFE, ele criptografa X com uma técnica de compartilhamento de segredo, dividindo-o em 3 partes (uma para cada servidor ABY^3). Cada servidor recebe duas das três partes, em um processo chamado *resharing*, para cada servidor realizar as operações necessárias. O hospital então faz o mesmo para o vetor de rótulos y relacionados a X.

Nos três servidores ABY^3, é possível treinar o modelo de forma privada com os dados criptografados e as operações definidas pelo protocolo. O vetor de pesos w é inicializado com valores zeros. Por fim, é obtido um modelo criptografado treinado nos dados criptografados, o qual fica repartido entre os três servidores.

Com o modelo treinado, o hospital pode requisitar predições privadas com entradas novas. Ele deve, primeiramente, criptografar a entrada nova e enviar para o servidores,

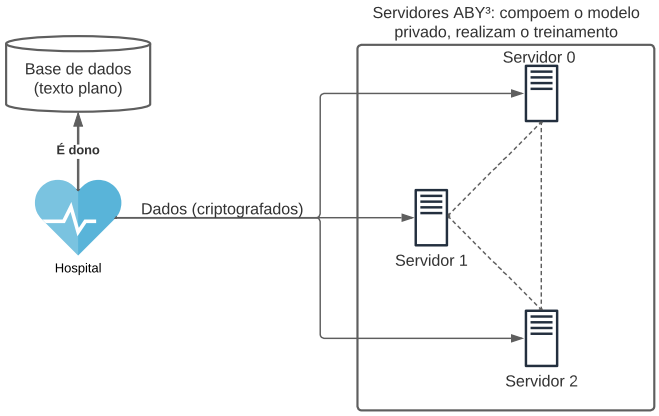

Figura 3 – Modelo de treinamento privado.

repetindo o processo do treinamento onde a entrada é dividida e compartilhada secretamente entre os três servidores. Os servidores então inserem a entrada no modelo treinado e devolvem uma resposta criptografada. Por fim, o hospital decifra a resposta dos servidores para obter o resultado coerente. A Figura 4 ilustra esse processo.

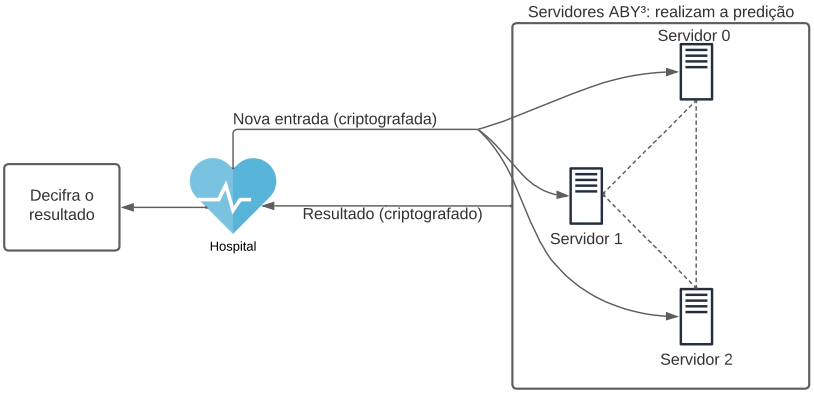

Figura 4 – Modelo de predição privada.

Como citado na Subseção 2.2.5, o ABY³ define um conjunto de operações que funcionam sobre um sistema de três servidores, os quais executam as operações privadamente utilizando diferentes técnicas de compartilhamento de segredo. Nele, para um valor $x \in Z_{2^k}$ ser compartilhado, ele é dividido em x_1, x_2 e x_3 de forma aleatória, de forma que $x_1 + x_2 + x_3 = x$. Em seguida, cada servidor recebe uma tupla dessas partes. O primeiro servidor recebe $\{x_1, x_2\}$, o segundo recebe $\{x_2, x_3\}$ e o terceiro recebe $\{x_3, x_1\}$

E assim, são definidas as operações que serão implementadas pelo TFE para reali-

zar o treinamento de um modelo. Sendo $[[x]]$ e $[[y]]$ valores privados no ambiente do ABY³, para realizar a adição os servidores calculam $[[x]]+[[y]] = [[x+y]] := (x1+y1, x2+y2, x3+y3)$ localmente. A multiplicação segura é mais complexa: sendo z_i o resultado parcial que cada servidor tem acesso, a operação de multiplicação é definida como:

$$\begin{aligned} z_1 &:= x_1y_1 + x_1y_2 + x_2y_1 + \alpha_1 \\ z_2 &:= x_2y_2 + x_2y_3 + x_3y_2 + \alpha_2 \\ z_3 &:= x_3y_3 + x_3y_1 + x_1y_3 + \alpha_3 \end{aligned} \tag{3.1}$$

O servidor i pode calcular z_i localmente, dadas suas partes de x e y. No entanto, é exigido que todos os servidores detenham duas das três partes (*resharing*). Para garantir isso, o protocolo especifica que o servidor i envie z_i para o servidor $i-1$. Os valores α são valores aleatórios $\in Z_{2^k}$ tais que $\alpha_1 + \alpha_2 + \alpha_3 = 0$. Eles são gerados por uma função pseudo-aleatória definida na fase de configuração dos servidores.

O cálculo da função sigmoide, necessária para alguns tipos de modelos de aprendizagem, é uma operação custosa na configuração de compartilhamento de segredo. Por isso, os autores do ABY³, inspirados pelo trabalho de Mohassel e Zhang [47], optam por utilizar um polinômio por partes, denominado como $\sigma(x)$.

$$\sigma(x) = \begin{cases} 0, & x < -0.5 \\ x + 0.5, & -0.5 \leq x < 0.5 \\ 1, & x \geq 0.5 \end{cases} \tag{3.2}$$

Porém, na implementação do protocolo feita pelo TFE, os autores observaram que a aproximação de três pontos do ABY³ resulta em uma precisão abaixo da média [46], e o substituem por uma solução de cinco partes, que se assemelha mais com a curva da função original. Essa solução é descrita na Equação 3.3, e na Figura 5 é possível observar que ela é mais semelhante a sigmoide do que o polinômio por partes que o ABY³ propõe.

$$\sigma(x) = \begin{cases} 10^{-4}, & x \leq -5 \\ 0.02776 \cdot x + 0.145, & -5 < x \leq -2.5 \\ 0.17 \cdot x + 0.5, & -2.5 < x \leq 2.5 \\ 0.02776 \cdot x + 0.85498, & 2.5 < x \leq 5 \\ 1 - 10^{-4}, & x > 5 \end{cases} \tag{3.3}$$

Assim, com o ambiente configurado e o conjunto de dados criptografado, é possível realizar o treinamento utilizando regressão logística como descrito na Equação 2.6 utilizando as operações citadas e substituindo o sigmoide pela função por partes exposta na Equação 3.3. O vetor de pesos w (o conjunto de pesos que se busca treinar iterativamente)

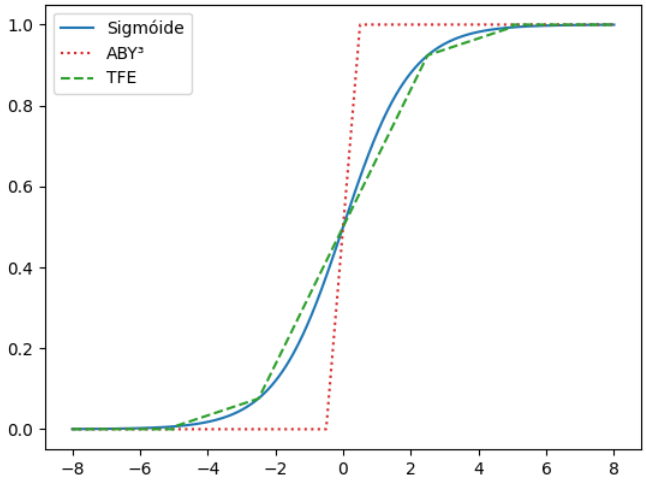

Figura 5 – Comparação entre a função sigmoide e as aproximações feitas pelo protocolo ABY³ e a biblioteca TFE.

é inicializado com valores iguais a zero. A parte de predição de valores novos consiste na operação de multiplicação do valor de entrada pelo vetor de pesos final.

3.2 Conjunto de dados

Neste trabalho, foram selecionados dois conjuntos de dados para evitar algum viés relacionado a um único conjunto de dados nos resultados obtidos no experimento. Ambos os conjuntos são de cunho médico e contém dados reais, se assemelhando aos tipos de dados que um hospital teria para realizar o treinamento de um modelo.

3.2.1 *Breast Cancer Wisconsin (Original)*

O conjunto de dados publicamente disponível *Breast Cancer Wisconsin (Original)* se encaixa na categoria de dados sensíveis, por ser extrapolado de exames de punção aspirativa por agulha fina de pacientes reais conduzido pelo Dr. William H. Wolberg no Hospital da Universidade de Wisconsin [48]. Os dados são preprocessados a partir das imagens resultantes do exame, sendo que o conjunto de dados tem o formato descrito na Tabela 2. Existem 699 elementos no conjunto, dos quais 241 (34,5%) são positivos e 458 (65,5%) são negativos.

O conjunto já contém dados pré-processados, então foi necessário tratar apenas 16 valores ausentes em 6291. Para corrigir isso, foi utilizada a biblioteca *pandas* para substituir estes valores pela mediana de sua respectiva coluna.

Depois, foram separados os valores de entrada dos rótulos. O vetor de rótulos

Tabela 2 – Formato do conjunto de dados *Breast Cancer Wisconsin (Original)*

Atributo	**Valores**
Clump thickness	1 a 10
Uniformity of cell size	1 a 10
Uniformity of cell shape	1 a 10
Marginal adhesion	1 a 10
Single epithelial cell size	1 a 10
Bare nuclei	1 a 10
Bland chromatin	1 a 10
Normal nucleoli	1 a 10
Mitoses	1 a 10
Class	2 (benigno) ou 4 (maligno)

consiste na última coluna da matriz original, facilitando a separação. Por fim, é aplicada uma função no vetor de rótulos, que utiliza os valores 2 e 4 como negativo e positivo, com o intuito de alterar estes valores para 0 e 1 respectivamente, alterando os valores para rótulos que são mais comuns de serem utilizados em treinamentos de modelos e até na programação em si (*false* e *true*).

3.2.2 *Pima Indians Diabetes Database*

Este conjunto de dados é disponibilizado pelo *National Institute of Diabetes and Digestive and Kidney Diseases*. Seu objetivo é prever se um paciente é ou não diabético, com base em certas medidas de diagnóstico incluídas no conjunto de dados [49]. Todos os pacientes aqui são mulheres com pelo menos 21 anos, descendentes do povo Pima. Os dados têm o formato descrito na Tabela 3. O conjunto contém 768 elementos, dos quais 268 (34,9%) são positivos e 500 (65,1%) são negativos.

Tabela 3 – Formato do conjunto *Pima Indians Diabetes Database*

Atributo	**Valores**
Pregancies	Inteiro
Glucose	Inteiro
Blood pressure	Inteiro
Skin thickness	Inteiro
Insuline	Inteiro
Body mass index	Ponto flutuante
Diabetes pedigree function	Ponto flutuante
Age	Inteiro
Outcome	1 (diabético) ou 0 (não diabético)

Este conjunto requer um pré-processamento mais extenso. Nele não existem valores ausentes, porém, existem valores zero nas colunas *Insulin, Glucose, Blood pressure, Skin thickness e BMI*. Para todas essas categorias, o valor zero é impossível. Como existem valores incorretos nessas colunas, porém em pouca quantidade, eles foram substituídos

pela mediana de suas classes. Já a coluna *Insulin* contém 48,69% de valores zero. Por ser incompleta demais, ela foi removida.

Por fim, o conjunto possui variáveis medidas em escalas diferentes, as quais não contribuem igualmente para o ajuste do modelo e podem acabar criando um viés [50]. Por isso, foi utilizado o *StandardScaler* da biblioteca *scikit-learn* para remover a média e dimensionar cada coluna de entrada para a variação da sua unidade.

3.3 Riscos do modelo

Mesmo utilizando proteção de privacidade, existem riscos associados ao uso do modelo no cenário proposto, os quais variam dependendo do nível de acesso que um adversário possui ao modelo. É importante ressaltar que o acesso ao classificador treinado deve ser controlado e, idealmente, deve ser utilizado internamente pela instituição para realizar novas predições, a fim de minimizar possíveis vulnerabilidades. No entanto, se o modelo for disponibilizado como serviço ou se um adversário obtiver acesso, a criptografia pode não ser suficiente para garantir a total proteção dos dados de treinamento. Mesmo com a utilização de técnicas criptográficas durante todo o processo, o simples acesso ao modelo pode revelar informações sobre o conjunto de dados de treinamento.

3.3.1 Adversário semi-honesto com acesso a um servidor

Um adversário semi-honesto é um atacante que segue o protocolo, mas pode tentar extrair informações do modelo treinado ou dos dados usados para treiná-lo. Para proteger os dados e o modelo contra ataques de adversários semi-honestos, são utilizadas técnicas de criptografia, como as empregadas na TFE.

A TFE utiliza técnicas de compartilhamento de segredo que exigem que os dados usados durante o processo de treinamento e predição sejam armazenados parcialmente em cada um dos três servidores envolvidos, garantindo que nenhum servidor tenha acesso completo aos dados. Dessa forma, mesmo que um servidor seja comprometido, o acesso ao seu conteúdo não revela nada sobre o conjunto de dados de treinamento ou qualquer outro dado utilizado durante o processo.

Essas medidas de segurança ajudam a proteger a privacidade dos dados e do modelo em ambientes onde há risco de ataques de adversários semi-honestos, permitindo que as organizações treinem modelos de aprendizado de máquina em dados confidenciais com segurança.

3.3.2 Adversário semi-honesto com acesso a mais de um servidor

O protocolo ABY3 e sua subsequente implementação pela TFE utilizam a técnica de *resharing* para proteger os dados durante o processo de compartilhamento de segredo.

Essa técnica divide os dados em três partes, mas cada servidor envolvido no processo de treinamento recebe apenas duas das três partes.

Embora essa técnica possa oferecer uma boa proteção contra adversários semi-honestos com acesso a apenas um servidor, é importante observar que ela não é totalmente segura contra o mesmo tipo de adversário com acesso a mais de um servidor. Isso ocorre porque, se um adversário semi-honesto obtiver acesso a pelo menos dois servidores distintos, ele poderá reconstruir o segredo e ter acesso aos dados em sua forma original.

Ao contrário dos adversários semi-honestos, um adversário desonesto é um invasor que não segue o protocolo e pode se comportar arbitrariamente. Ele pode tentar interromper o processo de treinamento ou manipular o modelo para atingir seus objetivos. Embora o protocolo ABY3 descreva técnicas para prevenir ataques de um adversário desonesto (através de operações mais complexas e custosas), a TFE não implementa essas medidas. Por isso, basta o acesso a apenas um servidor para que um adversário malicioso tenha acesso aos dados originais.

3.3.3 Adversário desonesto com acesso a um servidor

Diferentemente dos adversários semi-honestos, um adversário desonesto é um invasor que não segue o protocolo e pode agir de forma arbitrária. Ele pode tentar interromper o processo de treinamento, manipular o modelo para atingir seus objetivos ou obter informações sensíveis do modelo ou dos dados usados para treiná-lo.

Embora o protocolo ABY3 descreva técnicas para prevenir ataques de um adversário desonesto, na forma de operações mais complexas e custosas, a TFE não implementa essas medidas. Isso implica que, com acesso a apenas um servidor, um adversário malintencionado pode obter acesso aos dados originais. Essa obtenção ocorre pela combinação dos dados presentes em um servidor ao qual o adversário tem acesso com os dados requisitados de outro servidor, com o objetivo de reconstruir o segredo.

3.3.4 Adversário com acesso ao modelo

O acesso ao modelo de aprendizado de máquina pode permitir que um adversário execute ataques específicos, como descritos na Seção 2.2.2. Enquanto os ataques que exigem informações de gradiente (conhecidos como ataques de caixa branca) geralmente não representam uma ameaça, pois essa informação é privada no processo de treinamento, ataques de caixa preta podem ser uma ameaça sem a necessidade de alterar o ambiente de aprendizado do modelo. Se o adversário puder enviar consultas ao modelo e obter resultados, os métodos de caixa preta podem ser usados para realizar ataques e expor informações confidenciais presentes nos dados de treinamento [51].

3.3.5 Vazamento de dados

Uma vez que cada servidor recebe apenas uma parte gerada por um esquema de compartilhamento de segredos, um eventual vazamento de dados de um servidor durante o processamento desses dados não seria suficiente para revelar informações sobre o conjunto original de dados. Em teoria, seria necessário que ocorresse um vazamento em pelo menos dois servidores distintos para que um adversário pudesse reconstruir os dados originais.

3.4 Descrição dos experimentos

Para cada execução dos testes, quatro modelos de regressão logística foram criados. Para ambos os conjuntos de dados selecionados, foi treinado um modelo convencionalmente e outro utilizando as técnicas criptográficas disponibilizadas pela TFE, simulando o cenário descrito na Seção 3.1.

Para evitar o viés que um modelo fora da curva pudesse causar, foram realizadas 50 execuções para cada caso, com 20 épocas de treinamento. Para cada execução, o conjunto de dados foi dividido em 80% para treino e 20% para testes. Embora os tamanhos permaneçam os mesmos, em cada execução as entradas do conjunto de treino e de testes foram selecionadas aleatoriamente, permitindo uma validação cruzada. Por fim, o valor k utilizado pela TFE (número de bits utilizado para representar um valor numérico) foi definido como 128, seguindo os parâmetros utilizados pelos autores da TFE em um experimento similar de regressão logística [46]. O parâmetro de taxa de aprendizado foi definido como 0,01 para todos os casos. Para o conjunto de dados *BCW*, o tamanho do *batch* foi definido como 8 e para o conjunto *PID* foi escolhido o valor 64.

Após realizar o treinamento de cada modelo, o conjunto de testes foi utilizado para avaliar as métricas de precisão, revocação, *F1-score* e *Matthew's Correlation Coefficient*. As métricas analisadas são melhor detalhadas na Seção 3.5.

3.5 Análise dos resultados

Para avaliar os resultados, é possível gerar uma tabela de contingência chamada matriz de confusão de duas classes, representando quantos elementos foram previstos corretamente e quantos foram classificados incorretamente [52], como ilustra a Tabela 4. Assim é possível inserir esses valores em diferentes métricas para melhor avaliar os resultados.

Para representar a distribuição de resultados em cada caso foram utilizados gráficos de caixa. Este tipo de gráfico é composto por um retângulo que representa o intervalo entre o primeiro e o terceiro quartis (chamados Q1 e Q3), com uma linha vertical que indica a mediana, limites ilustrados na Figura 6. As extremidades da caixa (chamadas de bigodes)

Tabela 4 – Matriz de confusão.

		Valores obtidos	
		Positivos	Negativos
Valores reais	Positivos	*True positive (TP)*	*False negative (FN) (subestimação)*
	Negativos	*False positive (FP) (superestimação)*	*True negative (TN)*

estendem-se até os valores mínimo e máximo, exceto por valores que se encontram a uma distância maior que 1,5 vezes a amplitude interquartil (Q3 - Q1), sendo considerados possíveis *outliers* e indicados separadamente no gráfico.

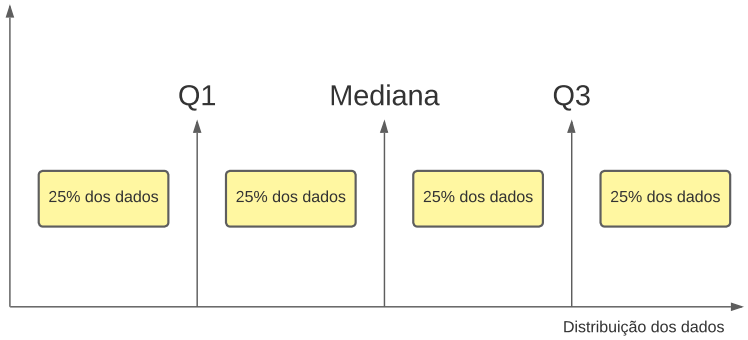

Figura 6 – Representação das delimitações Q1, mediana e Q3 em uma distribuição de dados.

3.5.1 Precisão

A precisão é uma medida estatística usada para determinar a proporção de casos positivos verdadeiros em uma população, dado um resultado de teste positivo. Contextualizando, ela é a probabilidade de que uma pessoa com teste positivo para uma determinada condição ou doença realmente tenha essa condição, ou doença. A fórmula da precisão é:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.4}$$

3.5.2 Revocação

A revocação é uma métrica que representa a proporção de casos positivos reais corretamente identificados como positivos. No contexto do problema modelado, a revocação é a porcentagem de pessoas com uma determinada condição que testam positivo para essa condição. A fórmula para calcular a revocação é a seguinte:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.5}$$

3.5.3 *F1-Score*

A F1-Score é a média harmônica entre a precisão e a revocação. Combinar essas duas métricas serve para analisar o desempenho geral do modelo e permite avaliar resultados mesmo em conjuntos desbalanceados (onde a distribuição de positivos e negativos é altamente discrepante). A F1-Score varia de 0 a 1, onde 1 é a melhor pontuação possível, indicando precisão e revocação perfeitas. Sua fórmula é:

$$\mathrm{F}_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{3.6}$$

3.5.4 *Matthew's Correlation Coefficient*

Para os modelos de classificação binária que foram testados neste trabalho, o *Matthew's Correlation Coefficient* (MCC) é uma taxa estatística mais confiável. Ele produz uma pontuação alta somente se os testes atingem bons resultados em todas as quatro categorias da matriz de confusão [53]. O resultado é proporcional ao tamanho dos elementos positivos e ao tamanho dos elementos negativos no conjunto de dados. O MCC varia de -1 a 1, com 1 indicando uma classificação perfeita, 0 indicando uma classificação aleatória e -1 indicando uma classificação completamente errada. Sua fórmula é:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

4 RESULTADOS

4.1 Precisão

Na Figura 7, observa-se que os resultados para o mesmo conjunto de dados considerando os cenários com e sem criptografia trazem resultados médios bem semelhantes. As execuções dos modelos treinado com o BCW criptografado produziram alguns resultados *outliers* a mais que sua contraparte sem criptografia. Além disso, as execuções dos modelos que utilizaram o PID criptografado tiveram uma precisão levemente pior que as execuções sem criptografia.

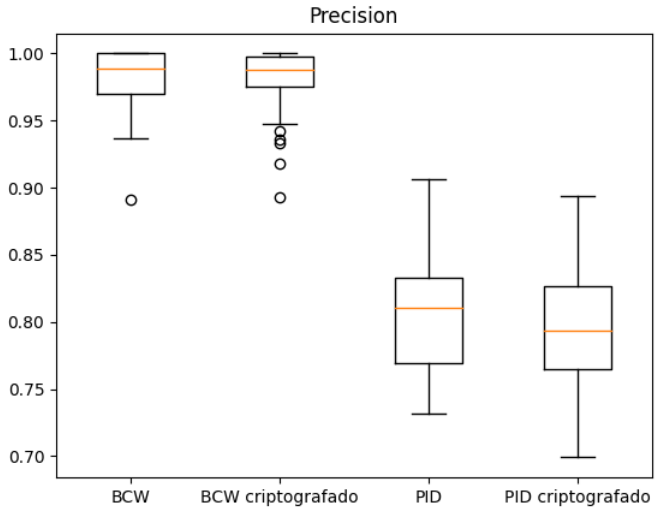

Figura 7 – Resultados obtidos na métrica de precisão.

4.2 Revocação

Como mostra a Figura 8, foram obtidos resultados semelhantes para o conjunto de dados BCW na maioria das vezes, sendo que o modelo convencional produziu mais *outliers*. Quanto ao PID, observamos uma quantidade considerável de resultados levemente melhores do modelo criptografado, porém, com medianas bem semelhantes.

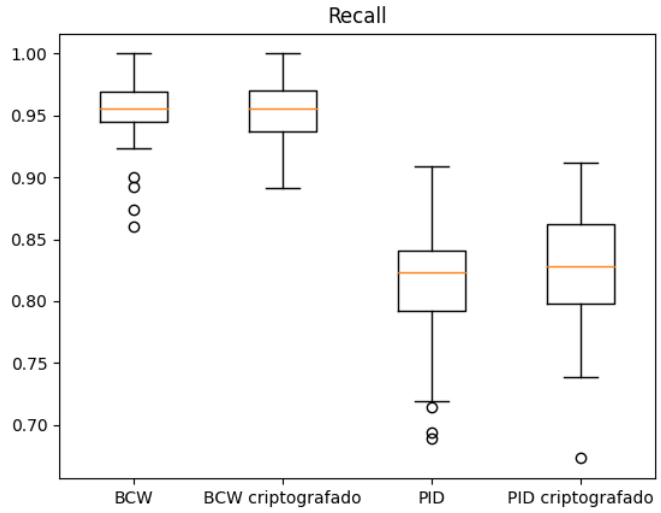

Figura 8 – Resultados obtidos na métrica de revocação.

4.3 *F1-Score*

Na Figura 9, é aparente que os resultados obtidos para essa métrica foram bem semelhantes para cada conjunto de dados. No BCW, os resultados coletados foram altamente semelhantes. Para o PID, os resultados variaram um pouco mais no modelo convencional, porém mais uma vez acusam medianas bastante semelhantes.

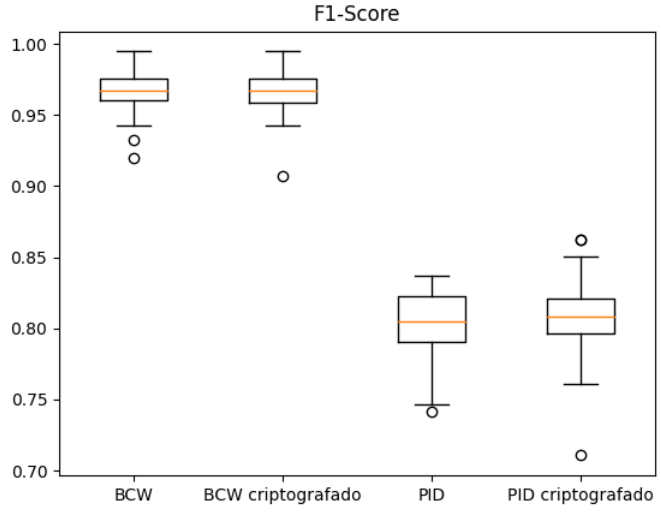

Figura 9 – Resultados obtidos na métrica *F1-Score*.

4.4 *Matthew's Correlation Coefficient*

A Figura 10 demonstra que os diferentes modelos apresentam resultados bastante semelhantes em cada conjunto de dados. Por se tratar de uma métrica recomendada para a avaliação geral da qualidade de um modelo [53], evidencia que, de forma geral, os modelos obtiveram resultados similares.

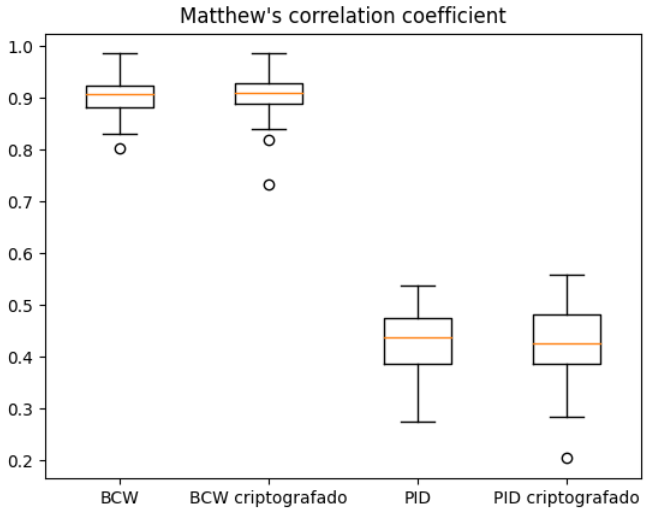

Figura 10 – Resultados obtidos na métrica *Matthew's Correlation Coefficient*.

5 CONCLUSÃO

Neste trabalho foi discutido o conceito de privacidade de dados na área da saúde, onde os dados médicos dos pacientes precisam ser tratados com o máximo de sigilo e segurança, devido às informações sensíveis que eles contêm. Foram destacadas as vantagens e riscos de se trabalhar com esses dados, onde, por um lado, a análise desses dados pode fornecer informações valiosas para a prevenção e tratamento de doenças, mas, por outro lado, o uso indevido dessas informações pode levar à violação da privacidade dos pacientes.

Foi apresentado um cenário em que uma instituição de saúde deseja treinar um modelo de classificação remotamente, utilizando os dados médicos dos pacientes, mas sem comprometer a privacidade desses dados. Para alcançar esse objetivo, foi proposto o uso de criptografia durante o treinamento do modelo, a fim de proteger o conjunto de dados de treinamento.

Para testar a viabilidade dessa proposta, foi implementada uma simulação do cenário, utilizando um esquema de três servidores e criptografia durante os processos. Foram treinados modelos de regressão logística utilizando o protocolo ABY^3 para realizar cálculos seguros. Os resultados coletados durante várias execuções mostraram que o modelo treinado utilizando criptografia não apresentou perda significativa na qualidade em comparação com modelos treinados sem criptografia.

Esses resultados são encorajadores por sugerirem que é possível treinar modelos de classificação em dados médicos sensíveis seguramente e preservando a privacidade dos pacientes, por meio do uso de técnicas de criptografia e de esquemas de múltiplos servidores. Essa abordagem pode ter implicações significativas para a pesquisa em saúde e para a melhoria dos cuidados médicos, garantindo, ao mesmo tempo, a privacidade e a segurança dos pacientes. Além disso, nada impede a utilização das técnicas estudadas em outras áreas que empregam aprendizado de máquina e preservação de privacidade.

Estudos futuros nessa área de pesquisa poderiam se concentrar em aprimorar ainda mais as técnicas de aprendizado de máquina com foco na preservação da privacidade não se limitando apenas a conjuntos de dados médicos. Um desafio importante é encontrar um equilíbrio entre a utilidade das informações obtidas e a proteção da privacidade dos dados. Além disso, seria interessante realizar estudos sobre a interpretabilidade e explicabilidade dos modelos, buscando compreender como as decisões são tomadas e fornecendo transparência aos profissionais de saúde.

REFERÊNCIAS

[1] SU, X. et al. Privacy as a service: Protecting the individual in healthcare data processing. *Computer*, IEEE, v. 49, n. 11, p. 49–59, 2016.

[2] KHAIRUNNAHAR, L. et al. Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked*, Elsevier, v. 16, p. 100189, 2019.

[3] LI, J. Ensuring privacy in a personal health record system. *Computer*, IEEE, v. 48, n. 2, p. 24–31, 2015.

[4] PERSON; MENN, J. *Exclusive Microsoft warns thousands of cloud customers of exposed databases*. Thomson Reuters, 2021. Disponível em: <https://www.reuters.com/technology/exclusive-microsoft-warns-thousands-cloud-customers-exposed-databases-emails-2021-08-26/>.

[5] SHAMIR, A. How to share a secret. *Communications of the ACM*, ACm New York, NY, USA, v. 22, n. 11, p. 612–613, 1979.

[6] GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. [S.l.]: O'Reilly Media, 2019.

[7] MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, p. 381–386, 2020.

[8] SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: IEEE. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. [S.l.], 2018. p. 1–6.

[9] BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4.

[10] CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: *Machine learning techniques for multimedia*. [S.l.]: Springer, 2008. p. 21–49.

[11] LIU, B.; LIU, B. Supervised learning. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, p. 63–132, 2011.

[12] DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: <http://archive.ics.uci.edu/ml>.

[13] DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: *Machine learning proceedings 1995*. [S.l.]: Elsevier, 1995. p. 194–202.

[14] HINTON, G.; SEJNOWSKI, T. J. *Unsupervised learning: foundations of neural computation*. [S.l.]: MIT press, 1999.

[15] BUHMANN, J.; KUHNEL, H. Unsupervised and supervised data clustering with competitive neural networks. In: IEEE. *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. [S.l.], 1992. v. 4, p. 796–801.

[16] TUCKER, A. B. *Computer science handbook*. [S.l.]: CRC press, 2004.

[17] CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, IEEE, v. 20, n. 3, p. 542–542, 2009.

[18] KAELBLING, L. P.; LITTMAN, M. L.; MOORE, A. W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, v. 4, p. 237–285, 1996.

[19] OTTERLO, M. V.; WIERING, M. Reinforcement learning and markov decision processes. *Reinforcement learning: State-of-the-art*, Springer, p. 3–42, 2012.

[20] WORSTER, A.; FAN, J.; ISMAILA, A. Understanding linear and logistic regression analyses. *Canadian Journal of Emergency Medicine*, Cambridge University Press, v. 9, n. 2, p. 111–113, 2007.

[21] MOLNAR, C. *Interpretable machine learning*. [S.l.]: Lulu. com, 2020.

[22] COHEN, J. E. What privacy is for. *Harvard law review*, JSTOR, v. 126, n. 7, p. 1904–1933, 2013.

[23] Brasil. *Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Redação dada pela Lei nº 13.853, de 2019. Brasília, DF: Senado Federal*. 2018. <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

[24] SOLOVE, D. J. I've got nothing to hide and other misunderstandings of privacy. *San Diego L. Rev.*, HeinOnline, v. 44, p. 745, 2007.

[25] NISSIM, K.; WOOD, A. Is privacy privacy? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society Publishing, v. 376, n. 2128, p. 20170358, 2018.

[26] GOSTIN, L. O.; JR, J. G. H.; VALDISERRI, R. O. Informational privacy and the public's health: the model state public health privacy act. *American Journal of Public Health*, American Public Health Association, v. 91, n. 9, p. 1388–1392, 2001.

[27] STARR, P. Smart technology, stunted policy: developing health information networks. *Health Affairs*, Project HOPE-The People-to-People Health Foundation, Inc., v. 16, n. 3, p. 91–105, 1997.

[28] VOGEL, S. *Tricare military beneficiaries being informed of stolen personal data*. WP Company, 2011. Disponível em: <https://www.washingtonpost.com/politics/tricare-military-beneficiaries-being-informed-of-stolen-personal-data/2011/11/23/gIQAcRNHtN_story.html>.

[29] STAFF, W.-N. *2 million affected by Shields Health Care Group cyberattack*. CBS Interactive, 2022. Disponível em: <https://www.cbsnews.com/boston/news/shields-health-care-group-data-breach-cyber-attack-massachusetts/>.

[30] TANKARD, C. What the gdpr means for businesses. *Network Security*, Elsevier, v. 2016, n. 6, p. 5–8, 2016.

[31] RIGAKI, M.; GARCIA, S. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*, 2020.

[32] BIGGIO, B.; ROLI, F. Wild patterns: Ten years after the rise of adversarial machine learning. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. [S.l.: s.n.], 2018. p. 2154–2156.

[33] NASR, M.; SHOKRI, R.; HOUMANSADR, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: IEEE. *2019 IEEE symposium on security and privacy (SP)*. [S.l.], 2019. p. 739–753.

[34] SHOKRI, R. et al. Membership inference attacks against machine learning models. In: IEEE. *2017 IEEE symposium on security and privacy (SP)*. [S.l.], 2017. p. 3–18.

[35] DWORK, C.; ROTH, A. et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, Now Publishers, Inc., v. 9, n. 3–4, p. 211–407, 2014.

[36] HE, Z.; ZHANG, T.; LEE, R. B. Model inversion attacks against collaborative inference. In: *Proceedings of the 35th Annual Computer Security Applications Conference*. [S.l.: s.n.], 2019. p. 148–162.

[37] ZHU, L.; LIU, Z.; HAN, S. Deep leakage from gradients. *Advances in neural information processing systems*, v. 32, 2019.

[38] ATENIESE, G. et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, Inderscience Publishers (IEL), v. 10, n. 3, p. 137–150, 2015.

[39] CATALANO, D. et al. Multiparty computation, an introduction. *Contemporary cryptology*, Springer, p. 41–87, 2005.

[40] BEAVER, D. Foundations of secure interactive computing. In: SPRINGER. *Advances in Cryptology—CRYPTO'91: Proceedings 11*. [S.l.], 1992. p. 377–391.

[41] BEIMEL, A. Secret-sharing schemes: A survey. In: SPRINGER. *Coding and Cryptology: Third International Workshop, IWCC 2011, Qingdao, China, May 30-June 3, 2011. Proceedings 3*. [S.l.], 2011. p. 11–46.

[42] BLAKLEY, G. R. Safeguarding cryptographic keys. In: IEEE COMPUTER SOCIETY. *Managing Requirements Knowledge, International Workshop on*. [S.l.], 1979. p. 313–313.

[43] KRAWCZYK, H. Secret sharing made short. In: SPRINGER. *Advances in Cryptology—CRYPTO'93: 13th Annual International Cryptology Conference Santa Barbara, California, USA August 22–26, 1993 Proceedings 13*. [S.l.], 1994. p. 136–146.

[44] MOHASSEL, P.; RINDAL, P. Aby3: A mixed protocol framework for machine learning. In: *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. [S.l.: s.n.], 2018. p. 35–52.

[45] YAO, A. C.-C. How to generate and exchange secrets. In: IEEE. *27th annual symposium on foundations of computer science (Sfcs 1986)*. [S.l.], 1986. p. 162–167.

[46] HONG, C. et al. Privacy-preserving collaborative machine learning on genomic data using tensorflow. In: *Proceedings of the ACM Turing Celebration Conference-China*. [S.l.: s.n.], 2020. p. 39–44.

[47] MOHASSEL, P.; ZHANG, Y. Secureml: A system for scalable privacy-preserving machine learning. In: IEEE. *2017 IEEE symposium on security and privacy (SP)*. [S.l.], 2017. p. 19–38.

[48] MANGASARIAN, O. L.; WOLBERG, W. H. *Cancer diagnosis via linear programming*. [S.l.], 1990.

[49] SMITH, J. W. et al. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the annual symposium on computer application in medical care*. [S.l.], 1988. p. 261.

[50] IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR. *International conference on machine learning*. [S.l.], 2015. p. 448–456.

[51] SHIN, J.; CHOI, S.-H.; CHOI, Y.-H. Is homomorphic encryption-based deep learning secure enough? *Sensors*, MDPI, v. 21, n. 23, p. 7806, 2021.

[52] CHICCO, D.; TÖTSCH, N.; JURMAN, G. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, BioMed Central, v. 14, n. 1, p. 1–22, 2021.

[53] CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, Springer, v. 21, p. 1–13, 2020.