CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Gestão de riscos para Lei Geral de Proteção de Dados

Risk management for General Data Protection Regulation

Gestión de riesgos para la Ley General de Protección de Datos

Leonardo Costa Farias[1], Bernardo da Eira Duarte[2], Karla Tereza Figueiredo Leite[3]

[1] Researcher, Telessaúde, UERJ, Rio de Janeiro (RJ), Brazil.
[2] Student, Instituto de Matemática e Estatística, UERJ, Rio de Janeiro (RJ), Brazil.
[3] Associate Professor, Telehealth and IME, UERJ, Rio de Janeiro (RJ), Brazil.

Corresponding Author: (M.Sc.) Leonardo Costa Farias
E-mail: leonardo.farias@uerj.br

Resumo

Objetivo: desenvolver novo modelo para avaliar riscos de privacidade relacionados à Lei Geral de Proteção de Dados (LGPD), em ambiente de saúde digital. Métodos: construção de modelo baseado em Lógica Fuzzy, considerando as boas práticas das normas técnicas da Associação Brasileira de Normas Técnicas (ABNT) e da International Organization for Standardization (ISO) para incorporar a incerteza no processo de avaliação assim como a explicabilidade dos resultados a partir da identificação de variáveis pertinentes ao conceito de risco para LGPD. Resultados: o novo modelo apresentou bons resultados quando comparados com outros modelos, além de, de forma diferenciada, incluir a explicação dos resultados obtidos. Conclusão: o sistema proposto utilizando o modelo intitulado Fuzzy-LGPD para Gestão de Riscos na Saúde Digital, apresentou resultados bastante promissores, possibilitando a identificação de riscos nos estudos de caso avaliados.

Descritores: Aplicação da Lei; Medição de Risco; Lógica Fuzzy.

Abstract

Objective: To develop a new model for assessing privacy risks related to the General Data Protection Law (LGPD) in the digital health environment. Methods: Construction of a model based on Fuzzy Logic, considering best practices from the technical standards of the Brazilian Association of Technical Standards (ABNT) and the International

Organization for Standardization (ISO) to incorporate uncertainty in the evaluation process as well as the explainability of results from the identification of variables relevant to the concept of risk for LGPD. Results: The new model showed good results compared to other models, and, differently, included the explanation of the obtained results. Conclusion: The proposed system using the model titled Fuzzy-LGPD for Risk Management in Digital Health presented very promising results, allowing the identification of risks in the evaluated case studies.

Descriptors: Law Enforcement; Risk Assessment; Fuzzy Logic.

Resumen

Objetivo: Desarrollar un nuevo modelo para evaluar los riesgos de privacidad relacionados con la Ley General de Protección de Datos (LGPD) en el entorno de la salud digital. Métodos: Construcción de un modelo basado en Lógica Difusa, considerando las mejores prácticas de las normas técnicas de la Asociación Brasileña de Normas Técnicas (ABNT) y de la Organización Internacional de Normalización (ISO) para incorporar la incertidumbre en el proceso de evaluación, así como la explicabilidad de los resultados a partir de la identificación de variables pertinentes al concepto de riesgo para LGPD. Resultados: El nuevo modelo mostró buenos resultados en comparación con otros modelos, y, de manera diferente, incluyó la explicación de los resultados obtenidos. Conclusión: El sistema propuesto utilizando el modelo titulado Fuzzy-LGPD para la Gestión de Riesgos en la Salud Digital presentó resultados muy prometedores, permitiendo la identificación de riesgos en los estudios de caso evaluados.

Descriptores: Aplicación de la Ley; Medición de Riesgo; Lógica Difusa.

Introduction

This study intersects multiple areas of knowledge, encompassing Medicine and other disciplines in the Health sector, with a special focus on Digital Health, as well as fundamental topics such as Privacy and Information Security, Risk Management, and Artificial Intelligence (AI), with an emphasis on Fuzzy Logic. The research also addresses legal and regulatory aspects, including compliance and adherence to technical standards established by the Brazilian Association of Technical Standards (ABNT) and the International Organization for Standardization (ISO), among others.

Digital health can expose personal data and sensitive personal data of patients, as well as of doctors and other professionals in this field, which places this type of scenario under the broad umbrella of the LGPD. Furthermore, the management of the risk of exposing these types of data should be routine in clinics, medical offices, and hospitals, which is not always done. During anamnesis, for example, numerous personal data and sensitive personal data are collected and consequently processed. Best practices for such processing must be adopted as a standard, not as an exception.

Digital Health represents the integration of information and communication technologies into the healthcare domain, aiming for continuous improvement in healthcare quality, increased operational efficiency, patient safety, and the promotion of health and general well-being. This field encompasses a broad spectrum of technological innovations, ranging from electronic health record systems and telemedicine platforms to mobile health monitoring applications and wearable devices for physical activity tracking. However, the implementation of this digital aspect faces significant challenges, including protecting user privacy and data security, ensuring accessibility, and suitability for use by different populations (PAUL, 2022) [1].

Privacy has been debated worldwide since the 19th century, with concepts evolving over decades. The rules protecting privacy give citizens around the world the ability to enforce their rights in the face of significant power imbalances, such as those held by large companies, notably Big Tech. These organizations are known for collecting and storing large amounts of Personally Identifiable Information (PII) from users of their products,

raising questions about privacy and data security. Due to their size and influence, these companies have been subject to regulatory scrutiny in many countries, with ongoing debates on how to address their business practices and impact on society at large (Privacy International, 2017) [2].

The enactment of the Brazilian General Data Protection Law (LGPD), Law No. 13,709, dated August 14, 2018, imposed a new wave of challenges and uncertainties on all public and private entities, as well as individuals processing PII, including the sensitive ones for commercial purposes. The risk of processing such data without adequate legal basis for processing, without a defined lifecycle, and without adherence to best practices - that is, in non-compliance with LGPD requirements - is substantially high, especially due to lack of regulatory compliance.

LGPD, in Articles 7 and 11, outlines a comprehensive set of guidelines for the management of PII. These articles stipulate the legal hypotheses for the processing of PII including the sensitive ones, respectively. The main objective of this regulation is to ensure privacy and guarantee the security of individuals' PII, establishing a set of principles, principles, and obligations that must be followed by all parties involved in the processing of such data (Brazil, 2018) [3].

The problem that motivated the development of this project is directly related to the processing of sensitive PII without adopting one of the legal hypotheses, or with the mistaken adoption, as provided for in LGPD, and compliance with good practices, principles, and other provisions provided for in the Law, which may result in applicable sanctions by the National Data Protection Authority (ANPD), lawsuits, and even sanctions by other public bodies (Luz, 2022) [4].

The motivation for this work arises from the growing need to manage privacy and protect personal and sensitive PII in an increasingly digitized and interconnected world. With the widespread adoption of technologies for Digital Health and the implementation of regulations such as LGPD, organizations must align their operations with new legal requirements and citizens' expectations regarding privacy and information security.

Risk Management using Fuzzy Logic is not something new and has already been widely addressed in academic literature. However, this work led to the development of an

innovative computational system that employs Fuzzy Logic for Risk Identification according to LGPD regulations in the context of Digital Health (SHUKRI, 2021) [6].

The remainder of this work is divided into three more sections. The Methods section presents a synthesized analysis of the adopted methodology, focusing on linguistic variables, both input and output, followed by the presentation of the digital product resulting from the aforementioned research. The Results section presents and discusses the obtained results, and finally, the last section provides a synthesis of the results so far, discussing their practical and theoretical implications and outlining perspectives for the continuation of the work.

Objective

The present study proposes the development of an innovative computational system that employs Fuzzy Logic for Risk Identification to comply with LGPD requirements in the context of Digital Health. This study explores both the benefits and challenges inherent in adopting this technology in society, also addressing the legal issues related to the implementation of technological solutions in the healthcare sector. To support this study, an extensive literature review was conducted to provide a critical and holistic analysis of the topic. It is expected that the findings of this research will significantly contribute to understanding the role of emerging technologies in the healthcare field and assist in formulating strategic decisions by professionals and entities in the sector, ensuring adherence to the requirements of this Law.

Methods

The methodology employed to achieve the objectives outlined in this study was based on an approach that aligned with the nature of the research objectives and provided detailed and contextualized insights into the management and treatment of personal and sensitive PII, in light of LGPD.

For the investigation of real cases, form links were sent to two private legal entities, consisting of two clinics: one specialized in medicine/telemedicine and the other in nutrition/teletnutrition. It is worth noting that the review by the Research Ethics Committee

(CEP) of the State University of Rio de Janeiro (UERJ) was waived, according to Resolution No. 510 of the National Health Council (CNS), dated April 7th, 2016. This decision was based on the nature of the collected data, which was strictly anonymized, limited to organizational information. Thus, no personal or sensitive PII were included, mitigating any potential risks that may exceed those encountered in everyday life, as defined in the mentioned Resolution. Article 1, paragraph VII - VII - research that aims at the theoretical deepening of situations that emerge spontaneously and contingently in professional practice, as long as they do not reveal data that may identify the subject.

These entities voluntarily participated in the evaluation, and their responsible collaborators answered, as legal entities, a questionnaire prepared by the researcher, providing the input information for the developed Fuzzy Inference System.

The Fuzzy Modeling developed to address the problem at hand was performed according to the Mamdani Inference Model, where both input and output variables are linguistic variables. The Mamdani model is more intuitive because its rules are expressed in natural language. This makes it easier for domain experts who may not have experience in mathematics or complex fuzzy systems to understand and build the fuzzy system. The minimal inference used in Mamdani is more straightforward to explain and justify in many contexts than the Larsen model. The output of the Mamdani system is a fuzzy set, which allows a direct interpretation of the rules and output of the system rather than the Tsukamoto model. Mamdani is also preferred when interpretability and ease of integration of human knowledge are more important than the mathematical precision offered by Takagi-Sugeno algorithms.

Nine linguistic variables were developed for the input variables, and only one, risk, was developed for the output variables. The choice of linguistic input variables was based on the Personal Data Inventory (IDP) model, made available by the Digital Government (Digital Government; Ministry of Management and Innovation in Public Services, 2021) [5]. The choice of the linguistic output variable was based on the Security and Privacy Risk Assessment Guide (Ministry of Economy, 2020) [7].

Considering personally identifiable information (PII) directly related to "dados pessoais" as defined in the LGPD, the following linguistic input variables were selected:

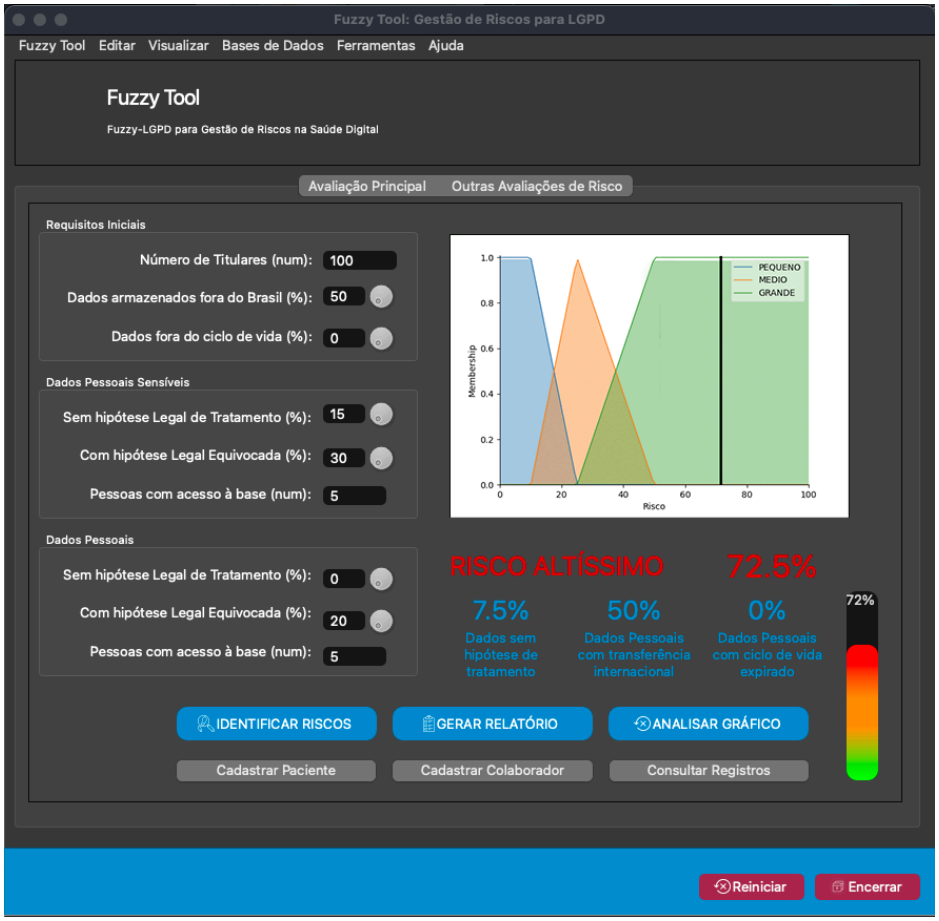
1. Number of data subjects (patients or collaborators);
2. Percentage of (PII) stored outside Brazilian territory;
3. Percentage of (PII) stored outside the legal lifecycle;
4. Percentage of sensitive (PII) without treatment hypothesis;
5. Percentage of sensitive (PII) with improper treatment hypothesis;
6. Number of people with access to sensitive (PII) ;
7. Percentage of (PII) without treatment hypothesis;
8. Percentage of (PII) with improper treatment hypothesis;
9. Number of people with access to (PII).

Only the classic AND and OR operations were employed to model the rules of the inference method, as proposed by Zadeh (1965) [8]. In total, 72 Fuzzy Inference Rules were generated (De Melo, 2020) [9].

Based on these linguistic variables, a desktop application with an interactive user interface (UI) was developed, referred to as a digital product, in the Python programming language using the NumPy, Scikit-Fuzzy, and Matplotlib libraries for conducting Fuzzy System analyses and generating graphics inherent to the digital product, which encompassed the proposition of a comprehensive identification of risks associated with the treatment of personal and sensitive PII, even in the absence of a legal treatment hypothesis established by LGPD. This proactive approach aims to anticipate and mitigate potential threats to data security and privacy, contributing to a more resilient and responsible approach to managing personal and sensitive PII in Digital Health.

On the application's front end, text input fields were provided for each linguistic input variable. After entering the data, clicking the "Identify risks" button returns the value of the linguistic output variable, i.e., risk. Figure 1 shows the main screen of the digital product.

Figure 1 - Fuzzy-LGPD Model for Risk Management in Digital Health



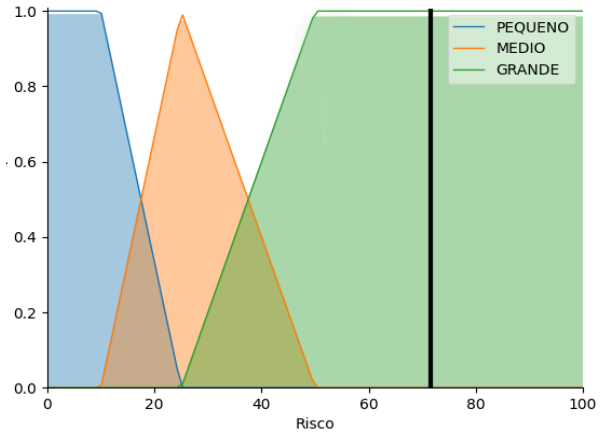
Considering x as the linguistic output variable, the risk was classified into the following ranges:

- For x ≤ 5: "Acceptable risk";
- For 5 < x ≤ 15: "Acceptable risk";
- For 15 < x ≤ 30: "Moderately low risk";
- For 30 < x ≤ 45: "Moderate risk";
- For 45 < x ≤ 60: "High risk";
- For 60 < x ≤ 85: "Very high risk";
- For 85 < x ≤ 100: "Catastrophic risk".

The graph presented in Figure 1, and in detail in Figure 2, illustrates the conversion of the outputs generated by the inference mechanism into precise values used by the control system. This is done using the technique known as defuzzification, employing the

centroid principle to calculate the final output value, i.e., the Risk. The centroid, represented by the vertical line, calculates the center of mass of the area under the Fuzzy output curve.

Figure 2 - Output inference variable: Risk



The centroid can be considered as the point along the x-axis where the fuzzy set balances, analogous to the center of mass of a physical object. The formula used to calculate the centroid is expressed according to equation (1) presented below, where μ(xi) represents the membership value for the point x in the universe of discourse:

$$\text{Centroid} = \frac{\sum_i \mu(x_i) x_i}{\sum_i \mu(x_i)} \qquad (1)$$

This mathematical expression is responsible for determining the weighted midpoint along the x-axis, considering the membership values associated with a fuzzy set. Defuzzification by centroid is a crucial tool in Fuzzy Logic, allowing the transformation of fuzzy sets into numerical values representing their central or average location. This technique has wide applicability in various fields, playing a significant role in problems involving uncertainty and imprecision.

Results and Discussion

Three cases were examined to assess the performance of the methodology and the Fuzzy-LGPD Model for Risk Management in Digital Health, for quantifying the risks

associated with the treatment of personal and sensitive PII by treatment agents in Digital Health: a. Comparison of the Digital Health System (DHS) model with the proposed model; b. Synthetic Test through the proposed model; and c. application of real cases to the proposed model.

No direct equivalences of the questions from the form were identified for their application to the DHS model. Regarding the synthetic test, the results were structured through individual tables for each linguistic input variable. The table headers indicate the corresponding variable, identified by the number of its respective linguistic input variable, following the previous presentation, as per the baseline alignment, which is aligned with the methodology step.

Additionally, a differentiation was implemented in the treatment of the output linguistic variable. In this regard, the variable was represented by the Obtained Risk (RO), to distinguish it from the Expected Risk (RE) reference, stipulated by experts in which the expected value was inserted during the test execution.

The values of RE were determined through the analysis of the test cases from the perspective of domain experts, with relevant experience in studies and work in the LGPD area, focusing on privacy, compliance, and data protection. The main purpose of these values was to provide a realistic interpretation of the expected responses from the model.

Table 1, presented below, highlights the variation of the parameter in question, Number of Subjects.

Table 1 – Parameter: Number of Data Subjects

1	2	3	4	5	6	7	8	9	RO	RE	|Δ| (%)
10^1	25	20	7	7	5	15	15	10	0,6	1,0	0,4
10^2	25	20	7	7	5	15	15	10	1,6	2,0	0,4
10^3	25	20	7	7	5	15	15	10	22,2	25,0	2,8
10^6	25	20	7	7	5	15	15	10	86,2	85,0	1,2
10^7	25	20	7	7	5	15	15	10	86,4	88,0	1,6
10^8	25	20	7	7	5	15	15	10	86,4	90,0	3,6

Regarding the responses from the research form, for theoretical deepening of the situation, arising from the treatments of PII, originating from the professional practices of the two clinics, medicine/telemedicine and nutrition/teletnutrition.

The universal sets of each clinic are small; even with input variables ranging from 0% to 60% for data stored outside the lifecycle, the output variable, risk, did not show significant variation. However, in empirical tests, a significant increase in the number of records implies a relevant increase in risk, their data were compared through Table 2, presented below.

Table 2 - Comparison of Online Form Results

Parameters	Nutrition Clinic	Medical Clinic
Number of Data Subjects (Patient or collaborator)	100	100
Percentage of data stored in the cloud	100%	100%
Percentage of data stored abroad	100%	100%
Percentage of data stored outside the lifecycle	0%	60%
Percentage of sensitive (PII) without hypothesis	0%	0%
Percentage of (PII) without hypothesis	0%	60%
Percentage of sensitive (PII) with wrong hypothesis	0%	0%
Percentage of (PII) with wrong hypothesis	30%	20%
Number of individuals with access to (PII)	2	1
Number of individuals with access to (PII)	1	1
Risk	20.1%	20.4%

It is noteworthy that although the percentage of data stored outside the lifecycle has a significant difference in the scenarios of the clinics, this did not cause a relevant impact on the calculated Risk, since the number of data subjects is considered small.

In the literature, there is no model that encompasses Fuzzy Logic, Digital Health, General Data Protection Law, Privacy, and Risk Management, which makes the proposal of this study innovative. The closest works in the literature to this model are the related works, compared to each other in Table 2 below:

Table 3 - Comparison of related works and the proposed model

Discipline	Garibaldi Model	SGD Model	Fuzzy-DP (Attaullah)	Harth Model	Proposed Model

LGPD	-	x	-	-	x
Digital Health	-	-	-	-	x
Fuzzy Logic	x	-	x	x	x
Risk Management	x	x	x	x	x
Privacy	-	x	x	x	x
GDPR	-	-	-	x	-

The comparison of models is subject to the limitations of the models' interpretations, as the variables of the models are not the same. Additionally, in crisp systems, a small change in the value of a variable can lead to a significant change in the output if a critical threshold is crossed. Fuzzy systems avoid this abrupt behavior, providing smooth transitions between different states. Fuzzy models are more robust to noise and small variations in input data, as Fuzzy Logic allows for continuous degrees of membership instead of binary decisions.

Conclusion

The aim of this study was to evaluate the feasibility of using a Fuzzy Inference System to assist in identifying the risk to the privacy of personal and sensitive PII of patients in a dedicated Digital Health system. The standard deviations and mean errors of the difference between the Obtained Risks and the Expected Risks demonstrated that the model behaved as expected, showing good performance and accuracy. These results suggest the continuation of exploring this topic from the perspective of Fuzzy Logic. The inference system proposed in this study was evaluated by two renowned experts in privacy and data protection, with a focus on the LGPD, reinforcing the relevance of this system.

Despite the diversity of tools available for compliance with the LGPD, the conclusion of this study highlights the scarcity of market applications based on management systems anchored in quantitative data and grounded in Fuzzy Logic, to replace boolean logic. Currently, applications are based on controls of international technical standards to ensure the reliability of their results. In this sense, it is recommended to incorporate Models that adopt Fuzzy Logic and encompass standardized management systems, such as the Quality Management System, Risk Management

System, Information Privacy Management System, Business Continuity Management System, and Information Security Management System, among others. This approach can significantly contribute to the effectiveness and robustness of the developed solutions.

Organizations, both software developers and those seeking compliance with the LGPD, must adopt best practices for the treatment of personal and sensitive PII and ensure respect for the rights of data subjects, including children and adolescents.

One advantage of the proposed model in this dissertation, the Fuzzy-LGPD Model for Risk Management in Digital Health, is that the explainability of models based on Machine Learning is increasingly demanded; it is more than a desirable requirement, since the answers need to be more easily understood by people. In this case, Fuzzy Logic meets this requirement naturally, i.e., intrinsically.

The implementation of a comprehensive system that encompasses the various stages of compliance with the LGPD can represent a significant investment. Solutions focused solely on risk, and based exclusively on probability, such as those offered in initiatives by Digital Government, often require specialization in risks and are not intuitively handled, in contrast to the solution proposed in this study. With the evolution of the proposed system, it is expected to enable Risk Management for personal and sensitive PII in digital health systems, dispensing with the need for the user to have expertise in risks or in the LGPD itself. The developed tool will provide a self-explanatory interface, allowing the user a clear understanding of the AI rules employed in the modeling.

Further research can be directed towards refining and improving the membership functions and Fuzzy rules used in the model. This may include integrating feedback from data protection and privacy experts and perspectives from end-users of digital health systems. The goal would be to improve the system's accuracy and efficiency in real-world situations, increasing its practical relevance and facilitating adoption by organizations.

The inclusion of other linguistic variables aims to improve the effectiveness of the proposed model, this time from boolean controls, such as those adopted by the DHS for its model. Other linguistic variables include the number of companies with which data is shared, the adoption of privacy management systems, information security, business continuity, among others, and the appointment of a Data Protection Officer (DPO).

Furthermore, the development of a more intuitive user interface and integration with other IT tools, such as electronic health record systems, could be explored to increase the usability and integration of the Risk Management System. This would help ensure that the solution can be easily implemented and used by professionals not specialized in risks or data protection legislation.

Finally, it is vital to explore the intersection between Fuzzy Logic and other emerging AI practices, such as deep learning and natural language processing, to develop even more advanced risk assessment systems. The inclusion of these technologies can provide a more dynamic and adaptable Risk Management System, capable of identifying and responding to new privacy threats in real-time.

By following these directions, it is expected to contribute to the continuous evolution of the field of privacy risk management and to the adaptation of organizations to the LGPD's requirements, promoting a safer and more reliable Digital Health ecosystem.

References

1. PAUL, M. Digitization of healthcare sector: A study on privacy and security concerns. https://www.sciencedirect.com/science/article/pii/S2405959523000243.

2. PRIVACY INTERNATIONAL. What Is Privacy? https://privacyinternational.org/explainer/ 56/what -privacy.

3. BRASIL. Lei n.º 13.709. https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709 compilado.htm.

4. LUZ, J. C. J. A Abordagem baseada no risco para a conformidade com a LGPD. https://www.conjur.com.br/2022-jan-05/jean-luz-abordagem-baseada-risco-conformidade-lgpd.

5. MINISTÉRIO DA GESTÃO E DA INOVAÇÃO EM SERVIÇOS PÚBLICOS; GOVERNO DIGITAL. Gestão de riscos. https://www.gov.br/governodigital/pt-br/seguranca-e-protecao-de-dados/ gestao-riscos.

6. SHUKRI, F. Mathematical Problems in Engineering - Experts' Judgment-Based Mamdani-Type Decision System for Risk Assessment. nov. 2021.

7. MINISTÉRIO DA ECONOMIA. Secretaria de Governo Digital. Guia de Avaliação de Riscos de Segurança e Privacidade. nov. 2020.

8. ZADEH, L. A. Fuzzy sets. Information and Control, v. 8, n. 3, p. 338–353, 1965.

9. DE MELO *et al*. Uma Avaliação das Medidas de Associação e Risco Fuzzy. https://jhi.sbis.org.br/ index.php/jhi-sbis/article/view/807/408.