Insper

TRIAGEM ONCOGENÉTICA MODELADA POR IA

Gabriel de Araujo Alves (gabrielaa3@al.insper.edu.br)
Keiya Nishio (keiyan@al.insper.edu.br)
Ricardo Mourão Rodrigues Filho (ricardomrf@al.insper.edu.br)
Sarah Azevedo Pimenta da Costa (sarahapc@al.insper.edu.br)

Trabalho de Conclusão de Curso

Relatório do Projeto Final de Engenharia
Versão Preliminar

São Paulo - SP – Brasil
FEVEREIRO 2024

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Gabriel de Araujo Alves
Keiya Nishio
Ricardo Mourão Rodrigues Filho
Sarah Azevedo Pimenta da Costa

TRIAGEM ONCOGENÉTICA MODELADA POR IA

Trabalho de Conclusão de Curso

Relatório do Projeto Final de Engenharia
Versão Preliminar

Relatório apresentado aos respectivos cursos de Engenharia (Computação/Mecânica/Mecatrônica) ou Ciência da Computação, como requisito parcial para a obtenção do título de Bacharel no respectivo curso.

Professor Orientador: Prof. Dr. Fábio José Ayres

Mentor:Dr. Jose Claudio Casali Da Rocha

Coordenador TCC/PFE: Prof. Dr. Luciano Pereira Soares

São Paulo - SP – Brasil
FEVEREIRO 2024

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Gabriel de Araujo Alves
Keiya Nishio
Ricardo Mourão Rodrigues Filho
Sarah Azevedo Pimenta da Costa

AI-MODELED ONCOGENETIC TRIAGE

Capstone Project Report
Preliminary Version

Undergraduate Capstone Project Report submitted to the respective (Computer/Mechanical/Mechatronic) Engineering programs or Computer Science program in partial fulfillment of the requirements for the Bachelor's Degree.

Advisor: Prof. Dr. Fábio José Ayres

Mentor: Dr. Jose Claudio Casali Da Rocha

General Coordinator TCC/PFE: Prof. Dr. Luciano Pereira Soares

São Paulo - SP – Brazil
FEBRUARY 2024

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Sumário

RESUMO **5**

ABSTRACT 6

1. INTRODUÇÃO **7**

1.1 ESCOPO DO PROJETO 8
1.2 RECURSOS 9
1.3 CRONOGRAMA 10
1.4 MAPEAMENTO DOS STAKEHOLDERS 11
1.5 RISCOS ENVOLVIDOS 13

REFERÊNCIAS 14

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

RESUMO

A Triagem Oncogenética por meio de questionário modelado por Inteligência Artificial (IA) representa uma inovação significativa no diagnóstico do câncer hereditário. Este estudo propõe uma abordagem que se baseia na análise criteriosa de parâmetros de história familiar para a seleção eficiente de pacientes aptos a realizar testes genéticos, sendo reconhecida como a estratégia mais custo-efetiva disponível atualmente. Com uma extensa base de dados de aproximadamente 10 mil pacientes atendidos e testados nos últimos quatro anos pelo A.C. Camargo, esta pesquisa se apoia em uma metodologia robusta, envolvendo a implementação de modelos probabilísticos e de inferência.

Um dos pontos-chave deste estudo é a possibilidade de utilização do pacote *PanelPro*, desenvolvido pelo laboratório *Bayesmendel* de Harvard, que oferece ferramentas avançadas para análise de dados genéticos. Além disso, o projeto se beneficia da expertise acumulada ao longo dos anos pelo A.C. Camargo em Oncogenética. Ao analisar uma amostra representativa de 4 mil pacientes, juntamente com os resultados de testes genéticos positivos e negativos, busca-se identificar os parâmetros que maximizam a sensibilidade, especificidade e a área sob a curva ROC (AUC), a fim de automatizar o processo de seleção de pacientes para testes genéticos.

A implementação bem-sucedida desta ferramenta tem o potencial de revolucionar a prática clínica, permitindo um encaminhamento mais eficiente e direcionado aos pacientes dos centros de referência (CRs) do A.C. Camargo para o Departamento de Oncogenética. Além disso, essa abordagem pode contribuir significativamente para a identificação precoce de indivíduos com predisposição genética ao câncer, possibilitando intervenções preventivas e terapêuticas mais eficazes. Em última análise, esse avanço tecnológico promete impactar positivamente a qualidade de vida e o prognóstico dos pacientes e suas famílias, representando um importante passo rumo à medicina personalizada e de precisão no combate ao câncer hereditário.

Palavras Chaves: Triagem Oncogenética, Inteligência Artificial, Testes Genéticos, Câncer Hereditário, História Familiar, Modelos Probabilísticos, PanelPro, Sensibilidade, Especificidade, Área sob a curva ROC (AUC).

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

ABSTRACT

Oncogenetic Screening through an Artificial Intelligence (AI)-modeled questionnaire represents a significant innovation in hereditary cancer diagnosis. This study proposes an approach that relies on thorough analysis of family history parameters for efficient selection of patients suitable for genetic testing, recognized as the most cost-effective strategy currently available. With an extensive database of approximately 10 thousand patients served and tested in the last four years by A.C. Camargo, this research is supported by a robust methodology involving the implementation of probabilistic and inference models.

One key point of this study is the possibility of utilizing the PanelPro package, developed by Harvard's Bayesmendel laboratory, which offers advanced tools for genetic data analysis. Furthermore, the project benefits from the expertise accumulated over the years by A.C. Camargo in Oncogenetics. By analyzing a representative sample of 4 thousand patients, along with the results of positive and negative genetic tests, the aim is to identify parameters that maximize sensitivity, specificity, and the area under the ROC curve (AUC), in order to automate the patient selection process for genetic testing.

The successful implementation of this tool has the potential to revolutionize clinical practice, enabling more efficient and targeted referral of patients from A.C. Camargo's reference centers (CRs) to the Oncogenetics Department. Additionally, this approach can significantly contribute to early identification of individuals with genetic predisposition to cancer, enabling more effective preventive and therapeutic interventions. Ultimately, this technological advancement promises to positively impact the quality of life and prognosis of patients and their families, representing a significant step towards personalized and precision medicine in the fight against hereditary cancer.

Keywords: Oncogenetic Screening, Artificial Intelligence, Genetic Testing, Hereditary Cancer, Family History, Probabilistic Models, PanelPro, Sensitivity, Specificity, Area under the ROC curve (AUC).

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

1. Introdução

No âmbito da medicina oncológica, o A.C. Camargo Cancer Center desempenha um papel crucial na detecção e tratamento precoces do câncer. Em sua missão de oferecer cuidados de saúde de excelência, a instituição tem investigado continuamente maneiras de aprimorar seus métodos de diagnóstico e triagem. Nos últimos quatro anos, o centro atendeu e testou cerca de 10 mil pacientes quanto à predisposição genética ao câncer, acumulando uma base de dados rica em informações sobre história familiar e resultados de testes genéticos.

Nesse contexto, surge o projeto de Triagem Oncogenética por meio de Inteligência Artificial, uma iniciativa voltada para o desenvolvimento de uma ferramenta que utiliza IA. O objetivo principal deste projeto é a criação de um sistema automatizado de triagem oncogenética, capaz de utilizar os dados disponíveis para identificar pacientes com maior probabilidade de apresentarem predisposição genética ao câncer.

A expectativa é que, com uma base de dados de história familiar de 4 mil pacientes, seja possível definir parâmetros relevantes para o desenvolvimento de um modelo de predição para ocorrência de genes em cada paciente. Isso permitirá uma triagem mais eficaz e direcionada, direcionando os pacientes mais adequados para a realização de testes genéticos, conforme os critérios estabelecidos.

Os benefícios potenciais deste projeto são significativos para o A.C. Camargo e seus pacientes. Além de melhorar a eficiência do processo de triagem, espera-se que a ferramenta proposta resulte em diagnósticos mais precoces, tratamentos mais eficazes e melhores resultados para os pacientes. Além disso, a automação desse processo pode liberar recursos e reduzir custos operacionais para a instituição.

Dada a vasta quantidade de dados disponíveis e a expertise do A.C. Camargo em oncologia e genética, o projeto é considerado viável. A implementação de uma ferramenta baseada em inteligência artificial para triagem oncogenética representa um avanço significativo na prática clínica e pode ser facilmente integrada aos processos existentes nos centros de referência do A.C. Camargo. Além disso, considerando o crescente interesse e investimento em tecnologias de saúde baseadas em IA, o projeto está alinhado com as tendências e demandas do mercado.

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

1.1 Escopo do projeto

- O que faz parte do Projeto:
    - Desenvolvimento de modelo de predição a fim de detectar possíveis predisposições genéticas.
    - Desenvolvimento de aplicativo web que recebe dados de triagem e aplica o modelo de predição desenvolvido de forma direta ao pesquisador (equipe médica).
    - Coleta, preparação e análise de dados de história familiar de pacientes atendidos pelo A.C. Camargo.
    - Definição de parâmetros relevantes para o modelo de predição.
    - Avaliação de impacto e monitoramento contínuo do desempenho da ferramenta.
- O que não faz parte do Projeto:
    - Desenvolvimento de testes genéticos em si.
    - Alterações estruturais nos sistemas de informação existentes do A.C. Camargo que não estejam relacionadas à integração da ferramenta de triagem.

O projeto enfrenta custos associados ao tempo dedicado à sua realização, além dos recursos humanos e, em alguma medida, da infraestrutura de TI necessários para o desenvolvimento da arquitetura que será proposta. Esses custos serão cuidadosamente monitorados e gerenciados, levando em consideração as restrições de tempo e orçamento estabelecidas. Quanto às restrições, o projeto está sujeito a limitações temporais e financeiras, exigindo uma gestão eficiente desses recursos para garantir o cumprimento dos objetivos dentro dos prazos e dos recursos disponíveis. Além disso, as premissas incluem o acesso total aos dados necessários para o desenvolvimento da ferramenta, bem como a estabilidade das políticas de privacidade e segurança de dados ao longo do projeto. Esses aspectos são fundamentais para garantir o progresso contínuo e bem-sucedido do projeto.

1.2 Recursos

Como um dos recursos o projeto pode se basear em um software (open source) de predição *PanelPro*, criado por Harvard, que se baseia em modelos preditivos para identificar membros da família com maior risco de mutações em genes de susceptibilidade ao câncer e

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

prever o risco de câncer associado a esses genes com base no histórico familiar. O *PanelPRO* é uma extensão do software anterior chamado *BayesMendel*, ambos com o mesmo propósito. Uma das vantagens do *PanelPro* é sua flexibilidade, permitindo uma ampla customização, incluindo a escolha dos genes e tipos de câncer a serem analisados. Além disso, o sistema oferece uma especificação de modelo personalizável e ajustes nos parâmetros conforme as necessidades do usuário, como frequências alélicas e penetrâncias. São fornecidas estimativas precisas de parâmetros padrão com base em uma extensa pesquisa na literatura científica, facilitando o seu uso. O software também permite a implementação de modificadores de risco para o câncer, como cirurgias profiláticas, raça e ancestralidade. Outra vantagem é a verificação abrangente dos dados de entrada do usuário, garantindo maior confiabilidade e precisão nas análises dos heredogramas (LIANG, 2022).

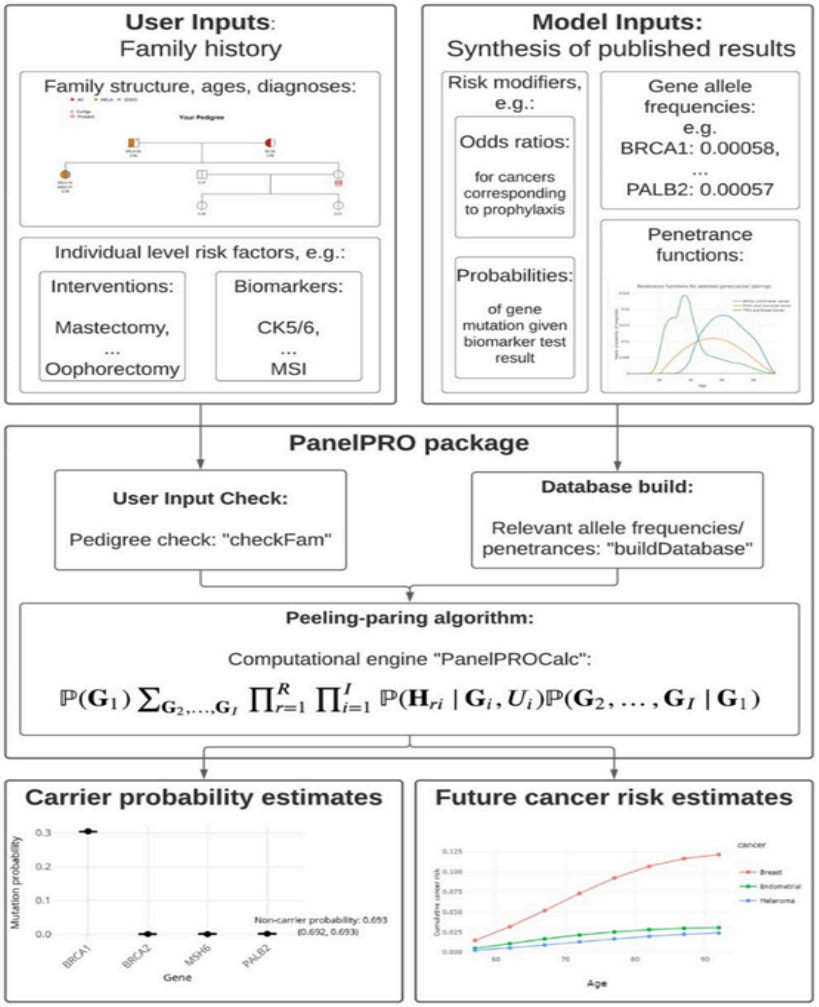

Figura 1 - PanelPro workflow

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

O software requer dois inputs: um fornecido pelo usuário e outro relacionado ao modelo. Os dados do usuário incluem informações sobre o histórico familiar e fatores de risco modificadores, enquanto os inputs do modelo são relacionados aos genes para prever futuras mutações.

Após receber os inputs, ocorre o pré-processamento para criar uma base de dados. Em seguida, o algoritmo, baseado nos métodos de "peeling" e "paring", processa os dados e prevê os riscos individuais. Por fim, o sistema apresenta os resultados, incluindo estimativas de portadores, intervalos de confiança das probabilidades de portadores com imputações para dados ausentes, previsões dos riscos futuros de câncer em intervalos de 5 anos e intervalos de confiança para essas previsões com imputações para dados ausentes.

1.3 Cronograma

Etapas	**JAN**	**FEV**	**MAR**	**ABR**	**MAI**	**JUN**
Relatório Preliminar		**X**				
Database e WebApp		**X**	**X**			
Dados Reais e construção modelo preditivo			**X**	**X**		
Seção de revisão relatório			**X**			
Relatório intermediário			**X**			
Treinamento modelo Preditivo			**X**	**X**	**X**	
Construção microsserviço de predição			**X**	**X**	**X**	
Interface CLI de consulta			**X**	**X**	**X**	
Relatório Final				**X**	**X**	

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Com os objetivos e metodologias definidos, é possível estimar os prazos de entrega do projeto, os quais devem ser baseados no calendário geral do PFE. Nesse sentido, há duas vertentes a serem consideradas: o fluxo de trabalho entre o relatório preliminar e intermediário e o fluxo de trabalho entre o relatório intermediário e final.

No que diz respeito ao primeiro fluxo de trabalho, espera-se que a equipe produza sprints semanais com o objetivo de gerar valor para a empresa concedente do projeto. Isso inclui apresentar um relatório consolidado e, se possível, realizar uma revisão de artigos considerados promissores para a pesquisa. Portanto, as demandas esperadas são: possuir um banco de dados para o treinamento do modelo, desenvolver um webapp que se conecte com o usuário e o banco de dados, e implementar um modelo que produza os resultados esperados.

Além disso, em relação ao segundo fluxo de trabalho, espera-se que a equipe tenha consolidado toda a infraestrutura esperada no projeto, seguindo as melhores práticas, como a utilização de containers e garantindo boa usabilidade para o usuário e o pesquisador. Também é importante começar a considerar iterações, revisar artigos de forma estruturada e preparar o relatório final para eventual publicação em revistas de saúde. Para alcançar esses objetivos, será essencial utilizar metodologias ágeis e realizar sprints semanais.

1.4 Mapeamento dos stakeholders

No contexto do presente relatório, é imprescindível destacar a importância do mapeamento dos stakeholders, uma ferramenta fundamental para a gestão estratégica de projetos. O mapeamento dos stakeholders envolve a identificação, análise e classificação das partes interessadas que podem influenciar ou ser influenciadas pelo projeto em questão. Esta abordagem permite uma compreensão abrangente do ambiente externo e interno, facilitando a tomada de decisões informadas e o desenvolvimento de estratégias de comunicação e engajamento eficazes. A seguir, apresenta-se uma tabela detalhada dos stakeholders identificados, na qual são evidenciados a posição de cada stakeholder no projeto, o papel desempenhado e as expectativas em relação ao projeto. Esta tabela serve como uma ferramenta visual para facilitar a análise e o planejamento estratégico das interações com cada parte interessada.

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Stakeholder	Posição	Papel no Projeto	Expectativa
Integrantes do grupo	Alto interesse e alta influência	Desenvolver modelo preditivo para a previsão de câncer	Entregar um modelo preditivo que cumpra integralmente com todas as especificações previamente estabelecidas.
Orientador	Alto interesse e alta influência	Direcionar o desenvolvimento do modelo e fornecer orientações técnicas e científicas	Entrega de um projeto bem-sucedido, bem como aprendizado dos alunos sobre projetos da indústria.
Equipe médica A. C. Camargo	Alto interesse e alta influência	Esclarecimento de questões técnicas relacionadas à medicina	Aplicar o modelo desenvolvido na prática clínica para a previsão de câncer
Equipe dados A. C. Camargo	Baixo interesse e alta influência	Disponibilizar dados necessários para o desenvolvimento e treinamento do modelo	Código claro, estrutura simples, automatizada e escalável.

Tabela 1 - Mapeamento dos Stakeholders

1.5 Riscos Envolvidos

Ao analisar os riscos envolvidos no projeto, a ética profissional no desenvolvimento do projeto, considerando o cliente pertencente ao sistema de saúde, emerge como um ponto crítico que requer atenção especial durante a construção do produto final. Considerando que lidamos com dados sensíveis relacionados à saúde dos pacientes, a violação da privacidade e confidencialidade desses dados pode representar um risco significativo. Isso pode ocorrer por meio de vazamentos de dados, acesso não autorizado por parte de terceiros ou até mesmo pela utilização inadequada dos dados coletados.

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Além disso, a manipulação inadequada dos dados de saúde dos pacientes pode resultar em consequências graves, como diagnósticos errôneos, tratamentos inadequados ou até mesmo danos à saúde dos pacientes. Portanto, é crucial garantir que todas as práticas de coleta, armazenamento, processamento e análise de dados sejam realizadas de acordo com os mais altos padrões éticos e regulatórios.

Sob essa perspectiva, é fundamental que os desenvolvedores do projeto estejam cientes e comprometidos com os princípios éticos para o desenvolvimento do projeto, garantindo a integridade, segurança e respeito aos direitos dos pacientes e do A.C Camargo. A negligência ou falta de ética profissional nesse contexto pode não apenas comprometer a qualidade do produto final, mas também ter sérias ramificações legais e éticas para todas as partes envolvidas.

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Referências

LEE, Gavin; LIANG, Jane W; ZHANG, Qing; HUANG, Theodore; CHOIRAT, Christine; PARMIGIANI, Giovanni; BRAUN, Danielle. *et al*. Multi-syndrome, multi-gene risk modeling for individuals with a family history of cancer with the novel R package PanelPRO. **Elife**, Cambridge, p. 1-9, ago. 2021. Disponível em: https://elifesciences.org/articles/68699. Acesso em: 20 fev. 24.

Clicksign fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

AC Camargo - Relatório final - Triagem Oncogenética - 2024.1.pdf

Documento número #fee3c84f-8b64-4c02-9b54-a5fc44ac5c00

Hash do documento original (SHA256): 14f8a0e7eed3b2f3b89ba3d41ea4056c7d90d6f3a0dc6ea30ff761fb0d1f5037

Assinaturas

Luciano Pereira Soares

CPF: 267.792.538-98

Assinou como testemunha em 01 jul 2024 às 13:14:45

Fabio Jose Ayres

CPF: 157.767.568-12

Assinou como parte em 04 jul 2024 às 10:03:52

José Cláudio Casali da Rocha

CPF: 779.612.777-49

Assinou como parte em 04 jul 2024 às 12:42:48

Log

01 jul 2024, 10:48:21	Operador com email JessicaAPS@insper.edu.br na Conta 22a9bded-c46a-42b4-a4e0-8d9a2a7f5083 criou este documento número fee3c84f-8b64-4c02-9b54-a5fc44ac5c00. Data limite para assinatura do documento: 31 de julho de 2024 (10:45). Finalização automática após a última assinatura: habilitada. Idioma: Português brasileiro.
01 jul 2024, 10:48:22	Operador com email JessicaAPS@insper.edu.br na Conta 22a9bded-c46a-42b4-a4e0-8d9a2a7f5083 adicionou à Lista de Assinatura: lpsoares@insper.edu.br para assinar como testemunha, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Luciano Pereira Soares.
01 jul 2024, 10:48:22	Operador com email JessicaAPS@insper.edu.br na Conta 22a9bded-c46a-42b4-a4e0-8d9a2a7f5083 adicionou à Lista de Assinatura: FabioJA@insper.edu.br para assinar como parte, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP. Dados informados pelo Operador para validação do signatário: nome completo Fabio Jose Ayres.
01 jul 2024, 10:48:22	Operador com email JessicaAPS@insper.edu.br na Conta 22a9bded-c46a-42b4-a4e0-8d9a2a7f5083 adicionou à Lista de Assinatura: casali.rocha@accamargo.org.br para assinar como parte, via E-mail, com os pontos de autenticação: Token via E-mail; Nome Completo; CPF; endereço de IP.

01 jul 2024, 13:14:45	Luciano Pereira Soares assinou como testemunha. Pontos de autenticação: Token via E-mail lpsoares@insper.edu.br. CPF informado: 267.792.538-98. IP: 130.126.255.243. Localização compartilhada pelo dispositivo eletrônico: latitude 40.1047552 e longitude -88.22784. URL para abrir a localização no mapa: https://app.clicksign.com/location. Componente de assinatura versão 1.900.2 disponibilizado em https://app.clicksign.com.
04 jul 2024, 10:03:52	Fabio Jose Ayres assinou como parte. Pontos de autenticação: Token via E-mail FabioJA@insper.edu.br. CPF informado: 157.767.568-12. IP: 189.121.200.1. Localização compartilhada pelo dispositivo eletrônico: latitude -23.5503616 e longitude -46.759936. URL para abrir a localização no mapa: https://app.clicksign.com/location. Componente de assinatura versão 1.906.0 disponibilizado em https://app.clicksign.com.
04 jul 2024, 12:42:48	José Cláudio Casali da Rocha assinou como parte. Pontos de autenticação: Token via E-mail casali.rocha@accamargo.org.br. CPF informado: 779.612.777-49. IP: 177.174.223.98. Componente de assinatura versão 1.906.0 disponibilizado em https://app.clicksign.com.
04 jul 2024, 12:42:49	Processo de assinatura finalizado automaticamente. Motivo: finalização automática após a última assinatura habilitada. Processo de assinatura concluído para o documento número fee3c84f-8b64-4c02-9b54-a5fc44ac5c00.

ICP Brasil

Documento assinado com validade jurídica.
Para conferir a validade, acesse https://www.clicksign.com/validador e utilize a senha gerada pelos signatários ou envie este arquivo em PDF.
As assinaturas digitais e eletrônicas têm validade jurídica prevista na Medida Provisória nº. 2200-2 / 2001

Este Log é exclusivo e deve ser considerado parte do documento nº fee3c84f-8b64-4c02-9b54-a5fc44ac5c00, com os efeitos prescritos nos Termos de Uso da Clicksign, disponível em www.clicksign.com.