



**NISAI
FACULDADE DE
MEDICINA - USP**



Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
30/01/2023	Henrique Godoy	1.0	Criação de Personas Criação dos Objetivos e Justificativas
31/01/2023	Raab lane	1.1	Criação do Value Proposition Canvas
31/01/2023	Guilherme Lima	1.2	Criação da Matriz de Risco
31/01/2023	Arthur Nisa	1.3	Criação da seção 4.1.3
31/01/2023	Arthur Tsukamoto	1.4	Criação da Introdução
01/02/2023	Henrique Godoy	1.5	Criação da Jornada do Usuário
01/01/2023	Arthur Tsukamoto	1.6	Criação da Compreensão do Problema
07/02/2023	Raab lane	1.7	Criação da Política de privacidade para o projeto de acordo com a LGPD
09/02/2023	Guilherme Lima	1.8	Criação das legendas das imagens
16/02/2023	Eduarda Gonzaga	2.1	Criação da seção 4.2
23/02/2023	Arthur Nisa	2.2	Atualização da seção 4.2
23/02/2023	Henrique Godoy	2.3	Atualização da seção 4.2
23/02/2023	Arthur Tsukamoto	2.4	Atualização da seção 4.2
23/02/2023	Guilherme Lima	2.5	Atualização da seção 4.2
23/02/2023	Raab lane	2.6	Atualização da seção 4.2

Sumário

1. Introdução 4

2. Objetivos e Justificativa 5

2.1. Objetivos 5

2.2. Proposta de Solução 5

2.3. Justificativa 5

3. Metodologia 6

4. Desenvolvimento e Resultados 7

4.1. Compreensão do Problema 7

4.1.1. Contexto da indústria 7

4.1.2. Análise SWOT 7

4.1.3. Planejamento Geral da Solução 7

4.1.4. Value Proposition Canvas 7

4.1.5. Matriz de Riscos 7

4.1.6. Personas 8

4.1.7. Jornadas do Usuário 8

4.2. Compreensão dos Dados 9

4.3. Preparação dos Dados e Modelagem 10

4.4. Comparação de Modelos 11

4.5. Avaliação 12

5. Conclusões e Recomendações 13

6. Referências	14
-----------------------	-----------

Anexos	15
---------------	-----------

1. Introdução

Nosso parceiro é o Instituto do Câncer do Estado de São Paulo (ICESP), pertencente à Faculdade de Medicina da Universidade de São Paulo (FMUSP), localizado próximo ao Hospital das Clínicas na Avenida Dr. Arnaldo, 251. Desde sua criação em 2008, o Instituto atua na área da oncologia, atendendo mais de 125 mil pacientes com foco no tratamento de diversos cânceres, tendo cerca de 5200 funcionários e prestadores de serviço e 500 leitos instalados.

Na área de Acreditações e Certificações, o ICESP recebeu a ONA (Organização Nacional de Acreditações) 1 e ONA 2 devido a sua forma de trabalho sistêmica e integrada. Além de certificações nacionais, o instituto também recebeu premiações internacionais como a Acreditação *Joint Commission International* (JCI) em 2014, 2017 e em 2020, além da *Commission on Accreditation of Rehabilitation Facilities* (CARF) em 2014.

Por fim, o ICESP possui um *Net Promoter Score* (Indicativo de Excelência) de 91, o que é considerado Excelente, ou seja, o padrão das áreas do Hospital, desde o atendimento até a realização dos exames possui uma nota excelente.

Anualmente, o ICESP trata diversos pacientes com diferentes tipos de câncer. Entretanto, o mais comum, principalmente, nas mulheres é o câncer de mama, o instituto trata entre 1000 a 1200 casos desse câncer, desde o estágio inicial até alguns casos de metástase (estágio mais avançado). Hoje em dia, existem 2 formas de tratamentos principais, sendo elas: a Terapia Adjuvante e a Terapia Neoadjuvante, porém com a evolução do câncer de mama a resposta desses tratamentos vem se tornando muito volátil.

Dessa forma, vem se estudando outros fatores que possam ajudar na melhor escolha do tratamento, dependendo das características do paciente.

2. Objetivos e Justificativa

2.1. Objetivos

O projeto tem como objetivo principal desenvolver um modelo preditivo para auxiliar o parceiro de negócios a encontrar a melhor solução para o problema da evolução do câncer. O modelo será utilizado para selecionar o tratamento de câncer mais eficiente para cada paciente, seja ele neoadjuvante (quimioterapia seguida de cirurgia) ou adjuvante (cirurgia seguida de terapia).

2.2. Proposta de Solução

A nossa solução propõe desenvolver um modelo preditivo para ajudar a escolher o melhor tratamento para cada cliente. Ele utilizará informações específicas sobre o paciente para identificar padrões e fazer previsões precisas sobre o tratamento mais eficaz.

2.3. Justificativa

A solução para o problema consiste em utilizar inteligência artificial para analisar dados de pacientes com câncer de mama e fornecer recomendações de tratamento personalizadas para cada caso. Dado o nível de complexidade envolvido nessa análise, é inviável para seres humanos realizá-la manualmente. O principal objetivo é entregar uma solução de alta qualidade, garantindo a precisão das recomendações de tratamento. Com isso, espera-se maximizar a eficácia dos tratamentos e, conseqüentemente, melhorar a qualidade de vida dos pacientes.

3. Metodologia

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

O ICESP está situado no setor da saúde pública focado no tratamento de pacientes com câncer.

O setor oncológico vem crescendo, principalmente no Brasil, devido ao aumento de casos de Câncer no País. De acordo com o Instituto Nacional de Câncer (INCA), entre 2023-2025 o Brasil deve ter cerca de 704 mil novos casos a cada ano, totalizando 2 milhões e 112 mil novos pacientes durante esse triênio.

Dentre esses 704 mil novos casos, cerca de 10% corresponde ao câncer de mama, o câncer de pele não melanoma retrata 30.4% dos casos, além do câncer de próstata (10,2%), pulmão (4.6%) e outros.

Com o crescimento da incidência de câncer, o investimento em infraestrutura, pesquisas e tratamento vem aumentando. Segundo o INCA, em 2018 foram investidos 3.4 bilhões de reais para o tratamento oncológico por meio do Sistema Único de Saúde (SUS), 41,1% ou 1.4 bilhões do investimento foi utilizado no tratamento de cânceres relacionados com o excesso de peso, como o câncer de mama. Além disso, o INCA espera que em 2030 sejam gastos 2.5 bilhões de reais apenas no tratamento dos cânceres de mama, colorretal e endométrio.

Portanto, é perceptível que o investimento na área da saúde oncológica vem aumentando, o que possibilita a utilização de novas tecnologias que possam ajudar esse setor como: Inteligência Artificial e *Machine Learning*.

Modelo do Negócio:

O ICESP é uma instituição do setor público, bancado e financiado pelo SUS, considerado um exemplo no tratamento de câncer, principalmente o câncer de mama. Atualmente, o instituto recebe pacientes de duas formas, uma parte dos pacientes vem do Hospital das Clínicas, enquanto que a outra parte é proveniente do sistema Central de Regulação de Ofertas de Serviços de Saúde (CROSS), criado pelo Estado de São Paulo em conjunto com instituições médicas que tem como objetivo organizar os recursos da saúde pública, direcionando os pacientes nos locais de tratamento corretos, dessa forma, caso o paciente tenha algum tipo de câncer ele será direcionado para algum centro oncológico ou instituição de combate ao câncer.

Principais Players:

Instituto do Câncer do Estado de São Paulo:

O ICESP possui tanto concorrentes públicos quanto privados, sendo eles: A rede Hebe Camargo, principalmente o Hospital Amaral Carvalho, o A.C.Camargo Cancer Center e o Centro Oncológico Família Dayan - Daycoval.

Rede Hebe Camargo(Hospital Amaral Carvalho):

A rede Hebe Camargo consiste em 76 hospitais que funcionam através do sistema CROSS, atendendo pacientes do estado de São Paulo. Dentre os hospitais, o Hospital mais renomado é o Amaral Carvalho localizado na cidade de Jaú, instituição que mais realiza transplante de medula óssea (cerca de 4000 transplantes em 2022), além de atender cerca de 70 mil pacientes com câncer anualmente.

Hospital A.C Camargo Cancer Center:

O Hospital A. C. Camargo Cancer Center, conhecido como o Hospital do Câncer fundado e inaugurado por Antônio Prudente foi um dos pioneiros no tratamento oncológico no Brasil na década de 50. Atualmente, possui mais de 5 mil funcionários e 6 unidades, sendo uma delas um Centro Internacional de Pesquisa(CIPE) responsável pelos estudos e pesquisas relacionadas aos diferentes tipos de cânceres.

Centro Oncológico Família Dayan - Daycoval:

O Centro Oncológico Família Dayan - Daycoval, pertencente ao Hospital Israelita Albert Einstein, se difere dos demais concorrentes, pois ele atende apenas pacientes do setor privado. Além disso, é a instituição com o maior número de Acreditações e Certificações do setor hospitalar e oncologia do Brasil, sendo o primeiro hospital brasileiro a receber a Acreditação *Joint Commission International* (JCI) e o único Centro oncológico brasileiro no top 20 dos melhores Hospitais Oncológicos do mundo pela pesquisa realizada pela *Newsweek*.

Tendências acerca do Modelo Preditivo na área da Saúde:

Atualmente, com o avanço tecnológico, a utilização de novas tecnologias como Inteligência Artificial e *Machine Learning* vem se tornando cada vez mais comum. Com a IA (Inteligência Artificial) e modelos matemáticos é possível criar um modelo preditivo, o qual utiliza-se de dados para prever resultados futuros. Dessa forma, podendo ser implementado em vários setores da sociedade, inclusive no setor da Saúde.

De acordo com estudos realizados pelos pesquisadores Michael McWilliams e Aaron L. da Universidade de Harvard, cerca de $\frac{3}{4}$ dos gastos do sistema de saúde americano é utilizado em apenas 17% dos pacientes. Isso ocorre, pois a maior parte dos pacientes possuem doenças crônicas, o que resulta em intervenções inoportunas e aumento de taxas de readmissão. Dessa

forma, o modelo preditivo poderia ser utilizado para prever algumas doenças crônicas, por exemplo: câncer de mama. Dessa forma, ao identificar essas doenças no estágio inicial ajudaria no tratamento do paciente, além de economizar dinheiro e aumentar a eficiência e produtividade dos hospitais.

Em suma, a utilização de modelos preditivos na área médica é bastante benéfico para o setor, já que a adoção desses modelos ajudariam médicos e equipes de suporte a tomarem a melhor decisão para determinado caso, além de facilitar na prevenção de custos adicionais ao identificar doenças crônicas em estágio inicial.

As 5 forças de Porter é um framework criado na década de 70 por Michael Porter, tal ferramenta permite realizar uma análise setorial a partir de 5 atores, mostrando como eles se relacionam com o negócio/produto.

As forças são:

- Poder de Negociação dos Fornecedores:

O Poder de Negociação dos Fornecedores em cima do Instituto do Câncer do Estado de São Paulo pode ser considerado alto, haja vista que por ser um hospital público, a sua verba é proveniente apenas do Estado de São Paulo. Dessa forma, o instituto tem uma verba limitada e ao negociar com os fornecedores os insumos para as operações, os fornecedores conseguem barganhar mais já que seus insumos médicos são importantíssimos para um hospital ou instituto médico.

- Poder de Barganha dos Clientes:

O Poder de Barganha dos Clientes é muito pequeno, pois, por se tratar de uma instituição pública e por exercer um serviço público, os clientes não conseguem barganhar uma diminuição do preço, já que é oferecido gratuitamente para todos.

- Ameaça de novos entrantes:

As principais ameaças são as instituições privadas, públicas e startups.

As instituições privadas possuem algumas vantagens em relação às instituições públicas, sendo elas: diversas formas de renda, ao invés de depender, unicamente, de apenas uma. Com uma quantidade maior de dinheiro, consegue fornecer uma qualidade de tratamento melhor ao cliente, além de uma infraestrutura bem montada, o que não acontece em algumas unidades de saúde pública.

O aumento de instituições públicas de saúde faz com que a verba voltada para a saúde seja repartida mais vezes. Com isso, o hospital acaba tendo menos dinheiro para comprar os insumos necessários ou oferecer um serviço melhor para os pacientes.

As Startups focam em inovações tecnológicas, isso faz com que ela esteja a frente dos outros concorrentes no quesito de tecnologia e métodos inovadores. Além disso, as startups possuem um grande investimento por meio de Venture Capitals e outros fundos, o que permite que ela continue desenvolvendo novas tecnologias, as quais demoram para chegar em hospitais públicos em que a verba é menor, tendo assim, uma vantagem em cima das instituições públicas e algumas privadas.

- **Ameaças de produtos substitutos:**

A Inteligência Artificial está crescendo no ramo da medicina. Em 2020-2021 uma pesquisa realizada pelo Instituto de Ciência e Tecnologia (ICT) da Universidade Federal de São Paulo em conjunto com o Instituto Tecnológico de Aeronáutica (ITA) utilizaram de técnicas de Machine Learning para criar um modelo preditivo para identificar os fatores de risco que levam pacientes infectado com Coronavírus a quadro graves e à internação.

Os cientistas de ambos os institutos utilizaram uma base de dados e por meio de modelos matemáticos identificaram padrões e tentaram prever resultados futuros. Esse mesmo modelo preditivo poderia ser convertido para identificar doenças crônicas, como o câncer de mama, o que seria uma ameaça de produto substituto.

- **Rivalidade entre concorrentes:**

O ICESP é uma instituição pública e sem fins lucrativos, que atua no tratamento de câncer. Apesar de não ter objetivos financeiros, a instituição compete com diversas outras entidades, tanto públicas quanto privadas, que oferecem serviços semelhantes e atendem ao mesmo público-alvo. Além disso, algumas metodologias e tratamentos exclusivos estão disponíveis apenas na rede privada. Dessa forma, é necessário que o ICESP se mantenha atualizado em relação às novas técnicas e tratamentos disponíveis, garantindo um atendimento de qualidade e um padrão de excelência no tratamento do câncer.

4.1.2. Análise SWOT

A matriz SWOT é uma ferramenta analítica que permite a avaliação dos pontos fortes e fracos de uma empresa, assim como as oportunidades e as ameaças que podem aparecer. Os quadrantes vermelhos são as forças internas e os lilás, as externas que podem impactar um negócio.

<p>Pontos Fortes</p> <ul style="list-style-type: none"> • Médicos e enfermeiros experientes. • Infraestrutura interna de qualidade. 	<p>Pontos Fracos</p> <ul style="list-style-type: none"> • Fila de espera extensa para atendimento médico. • Insuficiência de verbas para realizar alguns tratamentos. • Insumos médicos insuficientes para a demanda hospitalar.
<p>Oportunidades</p> <ul style="list-style-type: none"> • Credibilidade na área médica nacional e internacional. • Se relaciona diretamente à Faculdade de Medicina da USP, que ajuda na disponibilidade de mão de obra qualificada. 	<p>Ameaças</p> <ul style="list-style-type: none"> • Depende exclusivamente de verba do governo. • Não possuem a mesma verba para investimentos, comparado à hospitais particulares.

Imagem 1: Matriz SWOT do ICESP - USP.

4.1.3. Planejamento Geral da Solução

a) Descrição da solução a ser desenvolvida

Propomos uma solução personalizada de tratamento para o câncer de mama baseada em um modelo preditivo que leva em conta características do paciente, incluindo dados histopatológicos e demográficos. Dessa forma, esperamos maximizar a eficiência, minimizar os efeitos colaterais e aumentar a efetividade do tratamento de câncer de mama. Para garantir uma entrega de qualidade, tanto para o viabilizador do projeto quanto para o público-alvo, com o objetivo de melhorar a vida dos pacientes afetados pela doença.

b) Qual é o problema a ser resolvido

Alta variabilidade da resposta do câncer aos tratamentos convencionais.

c) Qual a solução proposta (visão de negócios)

Criar um modelo preditivo com base nos dados fornecidos pela FMUSP para indicar qual o melhor tratamento para cada paciente que foi diagnosticado com câncer de mama, aumentando o retorno gerado e diminuindo gastos com tratamentos incorretos.

d) Como a solução proposta deverá ser utilizada

Nossa solução será alimentada com dados pré-existentes dos pacientes do ICESP e que foram fornecidos pela FMUSP, para identificar padrões e fornecer informações valiosas para o corpo médico na tomada de decisões clínicas. A plataforma web será a interface para a inserção e acesso aos dados, tornando-a uma ferramenta de consulta fácil e acessível para todos os profissionais envolvidos no cuidado aos pacientes.

e) Quais os benefícios trazidos pela solução proposta

Através da análise de dados realizada pelo nosso modelo preditivo, o corpo médico terá acesso a uma previsão precisa sobre qual é o melhor tratamento para pacientes diagnosticados com câncer de mama, aumentando significativamente a probabilidade de cura. Além disso, o modelo irá levar em conta as condições específicas de cada paciente e indicar o tratamento mais adequado, o que irá maximizar a eficácia do tratamento.

f) Qual será o critério de sucesso e qual medida será utilizada para o avaliar

Diversos fatores serão considerados na análise, incluindo a taxa de sobrevivência do(a) paciente após o tratamento, quantos anos ele(a) viveu após o tratamento, comparação com previsões de taxa de sobrevivência baseadas em dados semelhantes, e previsão de quantos anos o(a) paciente pode ter de sobrevida. Essas métricas irão permitir que o modelo seja aprimorado e sua precisão aumentada, garantindo uma análise mais assertiva e confiável.

4.1.4. Value Proposition Canvas

O canvas da proposta de valor é uma ferramenta que permite relacionar o perfil do cliente com a proposta de valor, facilitando a compreensão de como a oferta é valiosa para o cliente. O canvas apresenta as dores e desejos do cliente do lado esquerdo, e os benefícios do produto do lado direito, demonstrando como o produto pode aliviar as dores e atender aos desejos do cliente.



Imagem 2: Value Proposition Canvas do produto.

Por meio dessa análise, é possível verificar que o produto produzido pelo grupo seria de um grande benefício para o ICESP, já que traria inúmeros ganhos para o cliente como citados na imagem acima.

4.1.5. Matriz de Riscos

A imagem a seguir representa uma tabela que ilustra os riscos e oportunidades que podem ser encontrados durante a confecção de uma ferramenta. Quanto mais ao meio da tabela, maior o impacto que esses riscos e oportunidades teriam no grupo. Já quanto mais acima, maior a probabilidade de ocorrência. Os riscos são representados por cores quentes, indicando maior preocupação, enquanto as oportunidades são representadas por cores frias, indicando maior benefício.


https://docs.google.com/spreadsheets/d/1COxFjLuMvimLm618_e1nsZao6N9EBQASCBj1ti0EUZ8/edit?usp=sharing

Riscos							Oportunidades						
Probabilidade	+90%	Pequenas dúvidas sobre o desempenho		Modelo não conseguir contemplar todas as expectativas dos integrantes			Integrantes adquirem mais conhecimentos sobre inteligência artificial e modelos preditivos		Modelo conseguir contemplar todas as expectativas do cliente			+90%	Probabilidade
	75%				Falta de dados suficientes		Pacientes conseguem ser tratados de uma maneira mais adequada	Redução do tempo de escolha de um tratamento				75%	
	50%			Falta de alinhamento dos integrantes sobre o tema	Médicos não se adaptam com o uso da ferramenta	Prazo final curto demais	Criação de empregos na manutenção da solução produzida	Aumentar o networking entre os membros do grupo				50%	
	25%		Bugs do Colab	Não entendimento de tópicos essenciais para a conexão do modelo	Modelo com falta de precisão	Falta de comunicação entre os integrantes	Aumentar a consciência sobre câncer de mama	Diminuição de custos				25%	
	10%	Ausência dos integrantes nas reuniões		Mal funcionamento dos computadores dos integrantes	Falta de apoio do corpo docente	Parceiro não gostar do produto final		Tratamento mais personalizado				10%	
		Muito Baixo	Baixo	Médio	Alto	Muito Alto	Muito Alto	Alto	Médio	Baixo	Muito Baixo		

Imagem 3: Matriz de risco do projeto.

4.1.6. Personas

Personas são representações personificadas do usuário típico do produto ou de outros indivíduos relacionados a ele. Esse método permite que o produto seja mais coerente com a realidade do público que o irá utilizar. Duas personas foram criadas para representar uma oncologista e uma paciente diagnosticada com câncer de mama. A perspectiva do paciente é importante para a equipe, já que ajuda a evitar recursos no produto que não levam em conta o paciente, sendo ele o foco do diagnóstico.



Sobre

Dr. Julia é formada em medicina pela Universidade Federal do Rio de Janeiro e tem especialização em oncologia. Ela trabalha em um hospital de referência em São Paulo e é responsável por cuidar de pacientes com câncer de mama.

Interesses

- Estar atualizada com as últimas pesquisas na área de oncologia.
- Procurando uma nova tecnologia que auxilie no tratamento de câncer.

Nome	Julia Chagas	<h3>Personalidade</h3> <ul style="list-style-type: none"> • Extrovertida • Intuitiva • Emocional • Criativa 	<h3>Dores</h3> <ul style="list-style-type: none"> • Indicação de tratamentos ineficientes. • Tratamentos não personalizados. • Gasto de tempo, atrapalhando a indicação de tratamentos. 	<h3>Necessidades</h3> <ul style="list-style-type: none"> • Mais tempo para poder cuidar de mais pacientes. • Melhorar a indicação de tratamentos nos casos de câncer de mama. • Tratamento de acordo com os dados do paciente.
Idade	42 anos			
Ocupação	Oncologista			
Localização	São Paulo, SP			
Educação	Mestrado			

Imagem 4 - Persona 1.


		Sobre Ana Paula, tem 54 anos. É uma mulher casada e mãe de dois filhos. Ela trabalha como gerente de RH em uma empresa de médio porte. Recentemente, foi diagnosticada com câncer de mama.	Interesses <ul style="list-style-type: none"> • Passar tempo com sua família e amigos. • Viagens. • Leitura. 											
<table border="1"> <tr><td>Nome</td><td>Ana Paula</td></tr> <tr><td>Idade</td><td>54 anos</td></tr> <tr><td>Ocupação</td><td>Gerente de RH</td></tr> <tr><td>Localização</td><td>São Paulo, SP</td></tr> <tr><td>Educação</td><td>Bacharelado</td></tr> </table>	Nome	Ana Paula	Idade	54 anos	Ocupação	Gerente de RH	Localização	São Paulo, SP	Educação	Bacharelado	Personalidade <ul style="list-style-type: none"> • Introvertida • Sensível • Social • Flexível 	Dores <ul style="list-style-type: none"> • Ela se frustra com o a incerteza. • Sente medo da falta de controle sobre sua saúde. • Erros médicos. 	Necessidades <ul style="list-style-type: none"> • Ela espera que o tratamento seja eficaz. • Não possua efeitos colaterais na sua qualidade de vida. • Médicos deem o melhor tratamento de acordo com seus dados. 	
Nome	Ana Paula													
Idade	54 anos													
Ocupação	Gerente de RH													
Localização	São Paulo, SP													
Educação	Bacharelado													

Imagem 5 - Persona 2.

4.1.7. Jornadas do Usuário

Julia Chagas			Necessidade:		
Cenário: Júlia é oncologista e está auxiliando no tratamento de uma paciente com câncer de mama.			Julia espera que o modelo consiga prever com eficiência o melhor tipo de tratamento para que a paciente consiga se recuperar do câncer.		
	Fase 1: Input de Dados	Fase 2: Execução do Modelo	Fase 3: Resultado	Fase 4: Análise dos Dados	Fase 5: Diagnóstico Final
Ações do usuário (Atividades)	Coleta e definição dos dados essenciais para execução do modelo Consulta dos registros médicos do paciente Preparação e organização dos dados para análise no modelo	Inserção dos dados coletados no modelo de análise Início do processo de análise de resultados pelo médico	Recebimento do melhor tipo de tratamento para o paciente	Análise detalhada da recomendação e utilização do modelo como um peso na decisão do melhor tratamento ao cliente	Comunicação do tipo de tratamento adequado ao paciente pelo médico
Oportunidades Desenvolver uma aplicação web que permita a captura de dados e melhore a experiência da Júlia. Criar um sistema de feedback, para que o sistema receba um feedback do médico sobre a escolha feita.			Responsabilidades Cabe ao time garantir uma boa assertividade quanto ao modelo preditivo no objetivo de auxiliar em aumentar a eficiência do médico em seu trabalho.		

Imagem 6 - Jornada de usuário com o produto.

4.1.8. Política de privacidade para o projeto de acordo com a LGPD

Somos uma equipe dedicada a proporcionar saúde e bem-estar por meio da tecnologia, desenvolvendo um modelo preditivo para que seja escolhido o melhor tratamento do câncer de mama levando em conta as informações únicas do paciente.

No âmbito da Lei Geral de Proteção de Dados Pessoais (LGPD), informamos que coletamos algumas informações pessoais, incluindo nome, idade, histórico médico, endereço e outras informações relevantes para o tratamento. Além disso, também coletamos dados não informados pelo usuário, como seu endereço de IP, localização e outros dados de navegação. Esses dados são coletados a partir da nossa plataforma eletrônica e são utilizados para fornecer o melhor modelo preditivo para a equipe médica, no objetivo de que o melhor tratamento de câncer de mama seja feito.

Os dados pessoais são armazenados em nossos servidores seguros e só serão mantidos por um período limitado, de acordo com nossa política de retenção de dados. Não utilizamos cookies ou tecnologias semelhantes. Compartilhamos os dados pessoais apenas com parceiros confiáveis e subcontratados que precisam desses dados para oferecer o melhor tratamento.

Adotamos medidas de segurança rigorosas para proteger os dados pessoais dos nossos pacientes, incluindo o acesso restrito aos dados. Seguimos rigorosamente as leis e regulamentos aplicáveis à proteção de dados pessoais.

Os pacientes têm o direito de solicitar e exercer seus direitos de proteção de dados, incluindo o direito de acesso, correção, exclusão e portabilidade de seus dados. Para exercê-los, basta entrar em contato conosco pelo e-mail inteli@inteli.edu.br.

O *Data Protection Officer* (DPO) da nossa equipe está sempre disponível para responder a quaisquer perguntas ou preocupações sobre o tratamento de dados pessoais.

Agradecemos a confiança depositada em nós e estamos comprometidos em proteger a privacidade e segurança dos dados pessoais de nossos pacientes.

4.2. Compreensão dos Dados

1. Exploração de dados:

O ICESP disponibilizou uma base de dados contendo 4 tabelas no formato de arquivo CSV (*Comma Separated Values*/valores separados por vírgula), formadas a partir de informações coletadas de prontuários médicos de pacientes com câncer de mama. As tabelas são: `df_demograficos`, `df_histopatologia`, `df_registro_tumo` e `df_pesoEaltura`, totalizando 77941 linhas e 103 colunas.

a) Cite quais são as colunas numéricas e categóricas.

As tabelas a seguir e suas respectivas categorizações estão em conformidade com o que foi escrito no código no Colab. Link: <https://github.com/2023M3T5-Inteli/grupo1>

`df_demograficos`

Numérico:

`record_id`

`tempo_seguinto`

idade_diagnostico
gestacao_idade
idade_primeira_menstruacao
tempo_amamentacao

Categórico:

escolaridade
sexo
raca
ultima informacão do paciente
ja_gravida
qual metodo?
consumo_alcool
historico_cancer
grau_parentesco_primeiro
grau_parentesco_segundo
grau_parentesco_terceiro
grau_parentesco_mama_primeiro_1_vez
grau_parentesco_mama_primeiro_mais_vezes
grau_parentesco_mama_segundo_1_vez
grau_parentesco_mama_segundo_mais_vezes
uso_anticoncepcional
tratamento
atividade_fisica
anti_her2_neoadjuvante

df_histopatologia

Numérico:

record_id

Categórico:

primeiro_diagnostico
grau_histologico
subtipo_tumoral
receptor_de_estrogenio
receptor_de_progesterona
ki67
her2ihc
her2fish

df_pesoEaltura

Numérico:

record_id
data
primeiro_peso
primeira_altura

primeiro_IMC
ultimo_peso
ultima_altura
ultimo_IMC
ultima_data
diferenca_peso
diferenca_tempo

df_registro_tumo

Numérico:

record_id
tempo_diagnostico
Data da primeira consulta institucional [dt_pci]
Data do diagnóstico

Categórico:

grupo_estadio_clinico
lateralidade_tumor
combinacao_tratamentos
classificacao_tnm_m
classificacao_tnm_n
classificacao_tnm_t_patologico
classificacao_tnm_n_patologico
recidiva_distancia
recidiva_local
recidiva_regional
estadio_clinico
cid_o
descricao_morfologia
descricao_topografia
morfologia_cid_o
metastase_cid1
metastase_cid2
metastase_cid3
metastase_cid4

3º Passo:

```
df_demograficos.describe()
```

	record_id	escolaridade	idade_diagnostico	sexo	raca	ultima_informacao_paciente	tempo_seguimento
count	4272.000000	3883.000000	4082.000000	4125.000000	221.000000	4270.000000	4270.000000
mean	48652.360487	3.286892	54.247801	0.008485	1.723982	2.108665	1475.003747
std	20659.519622	1.120554	13.574088	0.091733	1.232446	1.314601	859.622377
min	302.000000	0.000000	22.000000	0.000000	0.000000	0.000000	0.000000
25%	31013.000000	3.000000	45.000000	0.000000	1.000000	0.000000	956.250000
50%	53394.000000	4.000000	54.000000	0.000000	1.000000	3.000000	1282.000000
75%	65816.750000	4.000000	64.000000	0.000000	3.000000	3.000000	1817.750000
max	82240.000000	4.000000	88.000000	1.000000	3.000000	3.000000	4503.000000

Imagem 10: Resultado da função `describe()` aplicada na tabela `df_demograficos`.

Dessa maneira, por meio das informações obtidas das funções `info()` e `describe()`, foi possível identificar quais colunas estão mais preenchidas e poderiam trazer mais coesão para o modelo.

Na tabela `df_demograficos`, as colunas mais relevantes são:

- `record_id`: representa o identificador de cada paciente.
- `escolaridade`: mostra o nível de escolaridade do paciente.
- `idade_diagnostico`: idade do paciente no momento do diagnóstico.
- `ultima_informacao_paciente`: sinaliza a condição de vida do paciente (morto ou vivo).
- `tempo_seguimento`: quantidade de dias em que o paciente está seguindo o tratamento.
- `ja_gravida`: identifica se o paciente já engravidou.

Na tabela `df_pesoEaltura`, as colunas mais relevantes são:

- `record_id`: representa o identificador de cada paciente.
- `primeiro_peso`: peso do paciente na primeira consulta.
- `primeira_altura`: altura do paciente na primeira consulta.
- `primeiro_IMC`: IMC do paciente na primeira consulta.
- `ultimo_peso`: Peso do paciente na última consulta.
- `ultima_altura`: Altura do paciente na última consulta
- `ultimo_IMC`: IMC do paciente na última consulta.

Na tabela `df_histopatologia`, as colunas mais relevantes são:

- `record_id`: identifica o identificador de cada paciente.
- `primeiro_diagnostico`: identifica a região e o tipo do tumor.
- `grau_histologico`: identifica o estágio do tumor (I, II, IIA, IIB).
- `subtipo_tumoral`: verifica o subtipo do tumor (Luminal A, Luminal B, HER2 e Triplo negativo).
- `ki67`: mede o grau de proliferação do tumor.

Na tabela `df_registro_tumo`, as colunas mais relevantes são:

- `record_id`: identifica o identificador de cada paciente.

- `recidiva_local`: verifica a ocorrência de recidiva no mesmo local do tumor inicial.
- `recidiva_regional`: identifica a ocorrência de outro câncer próximo da região do tumor inicial.
- `recidiva_distancia`: sinaliza a presença de um outro câncer longe do local do primeiro tumor.

No primeiro conjunto de gráficos, é apresentada a quantidade de recidiva de acordo com o subtipo tumoral do paciente. O primeiro gráfico fornece uma visão geral, somando todos os subtipos e a quantidade de ocorrências de recidivas locais. É possível identificar que, dos 4177 pacientes, 3863 não apresentaram recidiva local, enquanto os 314 restantes tiveram o retorno do tumor.

Os gráficos seguintes demonstram a quantidade de recidiva local conforme o subtipo de tumor do paciente. São eles: Luminal A, Luminal B, Her2, Triplo negativo e, por fim, o metastático.

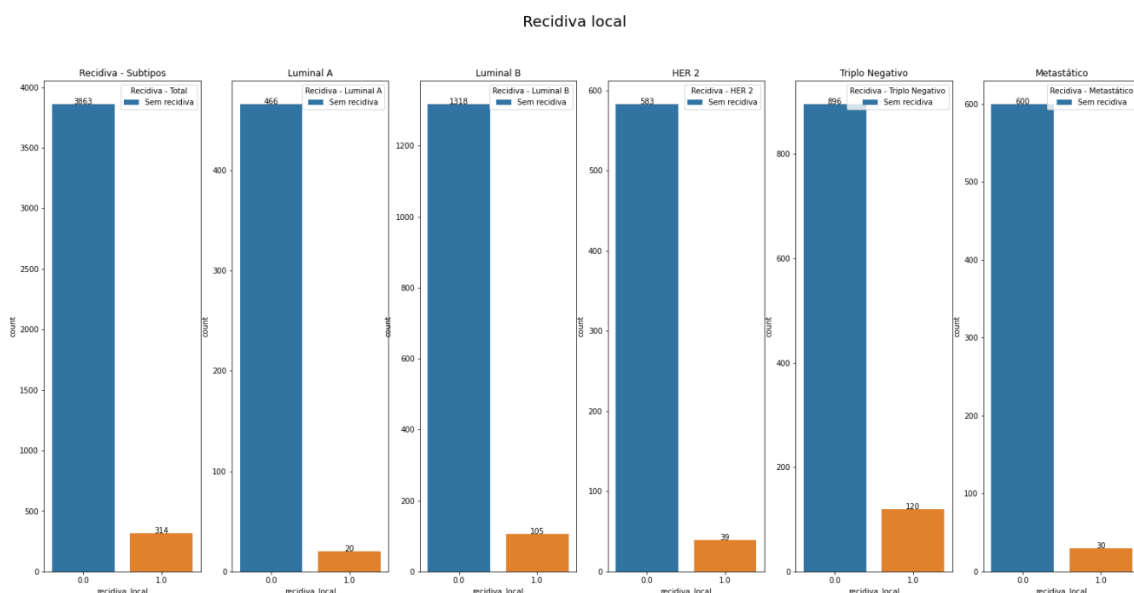


Imagem 11: Representação gráfica da Recidiva Local.

No segundo conjunto de gráficos, é demonstrada a relação entre recidivas regionais e a incidência desse tipo de recidiva de acordo com o subtipo do tumor. A diferença em relação ao primeiro gráfico apresentado é o tipo de recidiva escolhido para realizar a relação. Enquanto a recidiva local significa recorrência do tumor no mesmo local inicial, a recidiva regional representa a recorrência em uma região próxima do local do primeiro tumor. Dessa forma, é perceptível que, no caso da recidiva regional, dos 4177 pacientes, 3819 não apresentaram esse tipo de recidiva, enquanto os 258 restantes tiveram a recorrência do tumor.

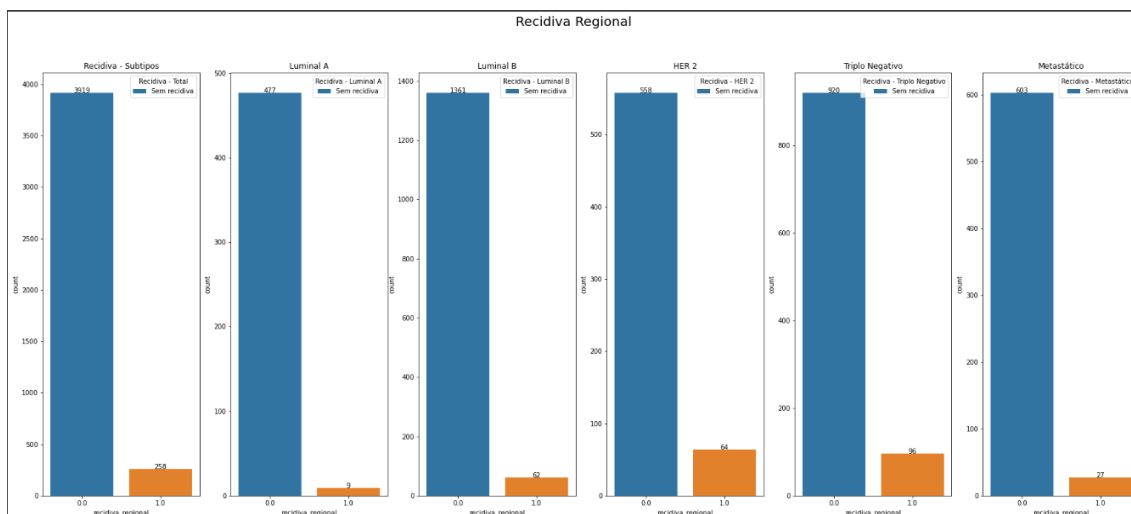


Imagem 12: Representação gráfica da Recidiva Regional.

No terceiro agrupamento de gráficos, é perceptível a relação entre recidivas à distância e o número de ocorrências gerais e de cada subtipo tumoral. A recidiva à distância se diferencia dos outros tipos de recidiva por ocorrer em uma região distante do tumor inicial.

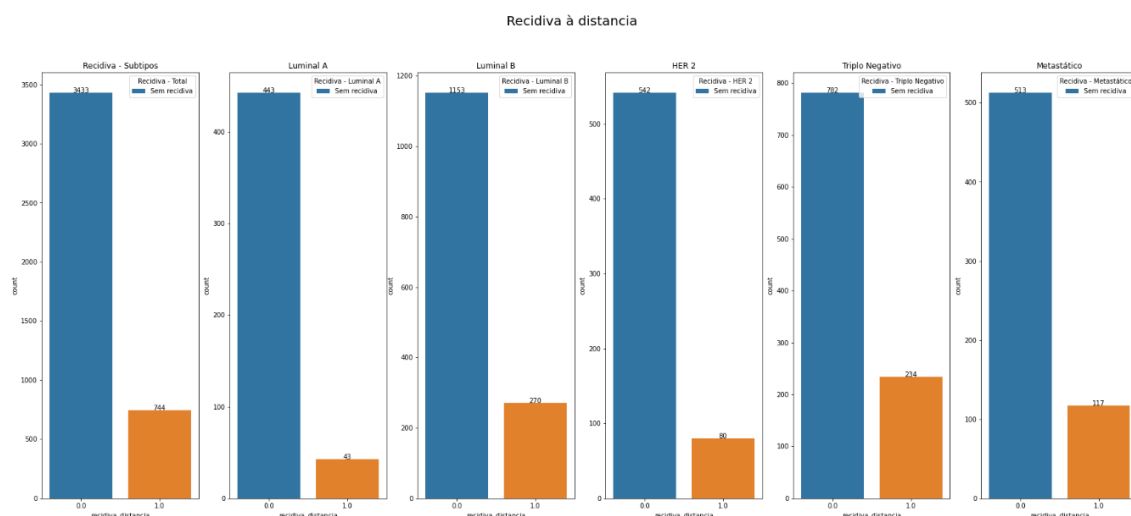


Imagem 13: Representação gráfica da Recidiva à distância.

2. Pré-processamento dos dados:

a) Cite quais são os outliers e qual correção será aplicada

As colunas 'escolaridade', 'idade_diagnostico', 'sexo', 'raca', 'gestacao_idade', 'tempo_amamentacao', 'idade_primeira_menstruacao', 'historico_cancer', 'grau_parentesco_primeiro', 'grau_parentesco_segundo', 'grau_parentesco_terceiro', 'tratamento', 'anti_her2_neoadjuvante', 'primeiro_diagnostico', 'grau_histologico', 'subtipo_tumoral', 'receptor_estrogenio', 'receptor_progesterona', 'ki67', 'her2_ihc',

'her2_fish', 'morfologia_cid_o', 'grupo_estadio_clinico',
 'classificacao_tnm_n', 'classificacao_tnm_m', 'metastase_cid1',
 'metastase_cid2', 'metastase_cid3', 'metastase_cid4',
 'combinacao_tratamentos', 'lateralidade_tumor', 'tempo_diagnostico',
 'classificacao_tnm_n_patologico', 'classificacao_tnm_t_patologico',
 'recidiva_distancia', 'recidiva_regional', 'recidiva_local', 'primeiro_peso',
 'primeira_altura', 'ultimo_IMC', 'diferenca_peso', 'coef_peso' e 'coef_peso_tempo'
 tiveram aplicada a correção de outliers, utilizando a função 'exclui_outliers'. Essa função remove as linhas do *DataFrame* que contêm valores considerados outliers em relação à coluna especificada, com base no cálculo do intervalo de valores aceitáveis para a coluna, que é definido como três vezes o desvio padrão da coluna em ambas as direções a partir da média da coluna. O *DataFrame* original é modificado inplace, sem retornar um novo *DataFrame*, de modo que, ao final do processo, as colunas especificadas não contenham mais outliers.

```
def exclui_outliers(df, col_name):
    intervalo = 3.4 * df[col_name].std()
    media = df[col_name].mean()
    lower_bound = media - intervalo
    upper_bound = media + intervalo
    df.drop(df[(df[col_name] < lower_bound) | (df[col_name] > upper_bound)].index, inplace=True)
```

Imagem 14: Utilização da função `exclui_outliers()`.

3. Hipóteses:

- a) Levantamento das três hipóteses com justificativa.

1. IMC afeta o tipo de tratamento escolhido

Depois de analisar os dados de vários pacientes e correlacionar seu IMC (Índice de massa corporal) com o tratamento recebido, percebemos que mulheres com obesidade e obesidade grave tendem a ser tratadas pelo método neoadjuvante. Para confirmar essa hipótese, realizamos uma análise das colunas relacionadas ao tratamento e IMC do paciente e plotamos ambos em gráficos usando o seguinte código:

```
import matplotlib.pyplot as plt

# Dividindo a figura em 1 linha e 5 colunas para os subplots
fig, axs = plt.subplots(1, 5, figsize=(15, 5))

# Definindo os dados e rótulos para cada faixa de IMC
magreza = merged_df.query("primeiro_IMC < 18.5")
normal = merged_df.query("primeiro_IMC < 24.9 and primeiro_IMC > 18.5")
sobrepeso = merged_df.query("primeiro_IMC < 29.9 and primeiro_IMC > 24.9")
obesidade = merged_df.query("primeiro_IMC < 39.9 and primeiro_IMC > 29.9")
obesidade_grave = merged_df.query("primeiro_IMC > 39.9")

counts = magreza['tratamento'].value_counts()
counts2 = normal['tratamento'].value_counts()
counts3 = sobrepeso['tratamento'].value_counts()
counts4 = obesidade['tratamento'].value_counts()
counts5 = obesidade_grave['tratamento'].value_counts()

# Plotando os gráficos de pizza em cada subplot
axs[0].pie(counts, labels=counts.index)
axs[0].set_title('Magreza')

axs[1].pie(counts2, labels=counts2.index)
axs[1].set_title('Normal')

axs[2].pie(counts3, labels=counts3.index)
axs[2].set_title('Sobrepeso')

axs[3].pie(counts4, labels=counts4.index)
axs[3].set_title('Obesidade')

axs[4].pie(counts5, labels=counts5.index)
axs[4].set_title('Obesidade Grave')

plt.show()
```

Imagem 15: Código desenvolvido pela equipe.



Imagem 16: Gráfico de Tratamento em relação ao IMC. Desenvolvido pela equipe.

A maior incidência do tratamento neoadjuvante ocorre devido às várias complicações que surgem quando a paciente apresenta algum grau de obesidade. De acordo com o estudo de 2019 intitulado "O Impacto da Obesidade no Diagnóstico e Tratamento do Câncer de Mama", essas complicações incluem aumento da probabilidade de desenvolver câncer de mama, maior chance de recorrência do câncer, maior número de complicações em cirurgias e menor

efetividade da quimioterapia sistêmica. Esses fatores contribuem para um grau mais grave de câncer, o que por sua vez tende a receber o tratamento neoadjuvante.

2. Nível de escolaridade não afeta a duração do acompanhamento

Com base na análise dos dados disponíveis, podemos formular a hipótese de que o nível de escolaridade não afeta a duração do acompanhamento de pacientes com câncer. Embora inicialmente tenha sido cogitado que pacientes com maior nível educacional teriam maior conscientização e, portanto, mais propensos a retornar para consultas médicas subsequentes, nossos resultados indicam o contrário.



Imagem 17: Gráfico de tempo de seguimento em relação à escolaridade. Elaborado pela equipe.

3. A influência da faixa etária no tempo de seguimento

Para testar essa hipótese, realizamos uma análise dos dados dos pacientes, correlacionando sua idade com o tempo de acompanhamento após o tratamento. Descobrimos que pacientes mais jovens tendem a ter um acompanhamento mais longo em comparação com pacientes mais velhos. Isso pode ser devido à maior preocupação das pacientes mais jovens em relação à recorrência do câncer, bem como a uma maior necessidade de monitoramento devido ao maior risco de desenvolver novos tumores.

Além disso, pacientes mais velhos podem estar mais suscetíveis a comorbidades, o que pode afetar sua capacidade de fazer consultas regulares e monitoramento de rotina. No entanto, é importante notar que essas tendências podem variar de acordo com o tipo específico de câncer de mama e suas características, bem como com outros fatores individuais dos pacientes, como histórico médico e estilo de vida.

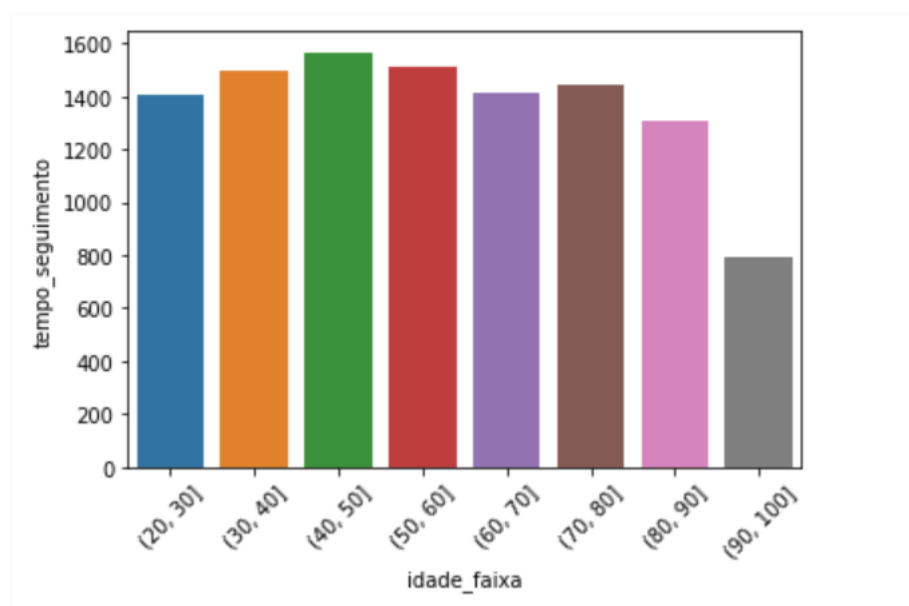


Imagem 18: Gráfico de tempo de seguimento em relação à idade. Elaborado pela equipe.

4.3. Preparação dos Dados e Modelagem

Caso seu projeto seja:

1. Modelo supervisionado:

- a) Modelagem para o problema (proposta de features com a explicação completa da linha de raciocínio).
- b) Métricas relacionadas ao modelo (conjunto de testes, pelo menos 3).
- c) Apresentar o primeiro modelo candidato, e uma discussão sobre os resultados deste modelo (discussão sobre as métricas para esse modelo candidato).

Caso seu projeto seja:

1. Modelo não-supervisionado:

- a) Modelagem para o problema (proposta de features com a explicação completa da linha de raciocínio).
- b) Primeiro modelo candidato para o problema.
- c) Justificativa para a definição do K do modelo.
- d) Escolha de um tipo de sistema de recomendação e a justificativa para essa escolha.

4.4. Comparação de Modelos

- *Escolha da métrica do modelo baseado no que é mais importante para o problema ao se medir a qualidade do modelo;*
- *Pelo menos três modelos candidatos com tuning de hiperparâmetros e suas respectivas métricas;*
- *Definição do modelo escolhido e justificativa.*

a) Escolha da métrica e justificativa.

b) Modelos otimizados.

- Apresentar três modelos e suas métricas.

- Os modelos apresentados foram otimizados utilizando algum algoritmo de otimização para os hiperparâmetros? Ex. Grid Search e Random Search.

c) Definição do modelo escolhido e justificativa.

4.5. Avaliação

Descreva a solução final de modelo preditivo e justifique a escolha. Alinhe sua justificativa com a Seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Descreva também um plano de contingência para os casos em que o modelo falhar em suas previsões.

Além disso, discuta sobre a explicabilidade do modelo e realize a verificação de aceitação ou refutação das hipóteses.

Se aplicável, utilize equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo e elaborar recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

6. Referências

ICESP, ICESP: Nossa História. 2022. Nossa História. Disponível em: <<https://icesp.org.br/nossa-historia/>>. Acesso em 31/01/2023.

ICESP, ICESP: Missão e Valores, 2022. Missão-visão-e-valores. Disponível em: <<https://icesp.org.br/missao-visao-e-valores/>>. Acesso em 31/01/2023.

ICESP, ICESP: Acreditações e Certificações, 2022. Acreditações e Certificações. Disponível em: <<https://icesp.org.br/acreditacoes-e-certificacoes/>>. Acesso em 31/01/2023.

ICESP, ICESP: NPS, 2022. Net Promoter Score. Disponível em: <<https://icesp.org.br/nps/>>. Acesso em 31/01/2023.

INCA estima 704 mil casos de câncer por ano no Brasil até 2025. gov.br, 2022. Disponível em: <<https://www.gov.br/inca/pt-br/assuntos/noticias/2022/inca-estima-704-mil-casos-de-cancer-por-ano-no-brasil-ate-2025>>. Acesso em 31/01/2023.

A.C.Camargo, ACC: Nossa História. 2022. Nossa História. Disponível em: <<https://accamargo.org.br/institucional/nossa-historia>>. Acesso em 31/01/2023.

Amaral Carvalho, Amaral Carvalho: Fundação. 2022. Fundação. Disponível em: <<https://amaralcarvalho.org.br/fundacao/sobre>>. Acesso em 31/01/2023.

Einstein, Einstein: Sobre. 2022. Quem Somos. Disponível em: <<https://www.einstein.br/sobre-einstein>>. Acesso em 31/01/2023.

Einstein, Einstein: Acreditações e Certificações, 2022. Acreditações e Certificações. Disponível: <<https://www.einstein.br/sobre-einstein/qualidade-seguranca/acreditacoes-certificacoes-designacoes>>. Acesso em 31/01/2023.

Gastos do SUS com cânceres que poderiam ser prevenidos com atividade física chegarão a 2.5 bilhões em 2030. gov.br, 2022. Disponível em: <<https://www.gov.br/inca/pt-br/assuntos/noticias/2022/gastos-do-sus-com-canceres-que-poderiam-ser-prevenidos-com-atividade-fisica-chegarao-a-r-2-5-bilhoes-em-2030>>. Acesso em 31/01/2023.

Gastos do SUS com cânceres associados ao excesso de peso somam 41.1% do investimento em tratamento oncológico. Instituto Nacional do Câncer, 2021. Disponível em: <<https://www.inca.gov.br/noticias/gastos-do-sus-com-canceres-associados-ao-excesso-de-peso-somam-41-do-investimento-em>>. Acesso em 31/01/2023.

Análise Preditiva: como ela impacta o setor de saúde. MedSimples, 2020. Disponível em: <<https://www.medsimples.com.br/medtech/analise-preditiva/>>. Acesso em 31/01/2023.

Modelagem preditiva aumenta eficiência de sistemas de saúde.
 Estadão, 2021.

Disponível

em: < <https://summitsaude.estadao.com.br/saude-humanizada/modelagem-preditiva-aumenta-eficiencia-de-sistemas-de-saude/> > . Acesso em 31/01/2023.

Como funciona a Central de Regulação de Ofertas de Serviços de Saúde - CROSS. Santa Casa, 2020.

Disponível

em:

< <https://santacasape.com.br/site/2020/06/19/como-funciona-a-central-de-regulacao-de-ofertas-de-servicos-de-saude-cross/> > . Acesso em 31/01/2023.

"O Impacto da Obesidade no Diagnóstico e Tratamento do Câncer de Mama", 2019. Disponível

em: < <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6437123/> > . Acesso em 23/2/2023.

Anexos