



**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
09/03/2023	Henrique Godoy	3.1	Criação das seções 3 e 4.3
09/03/2023	Lucas Galvão	3.2	Atualização das seções 3 e 4.3

09/03/2023	Guilherme Lima	3.3	Atualização das seções 3 e 4.3
09/03/2023	Arthur Tsukamoto	3.4	Atualização das seções 3 e 4.3
09/03/2023	Raab lane	3.5	Atualização das seções 3 e 4.3
23/03/2023	Guilherme Lima	4.1	Criação da seção 4.4
23/03/2023	Arthur Tsukamoto	4.2	Atualização da seção 4.4
23/03/2023	Raab lane	4.3	Atualização da seção 4.4

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 ICESSP 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 ICESSP 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 ICESSP 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

CRISP DM

Cross Industry Standard Process [for] Data Mining: É uma metodologia com abordagem extremamente eficiente utilizada na resolução de problemas e projetos envolvendo dados, ramificada em 6 estágios ou etapas:

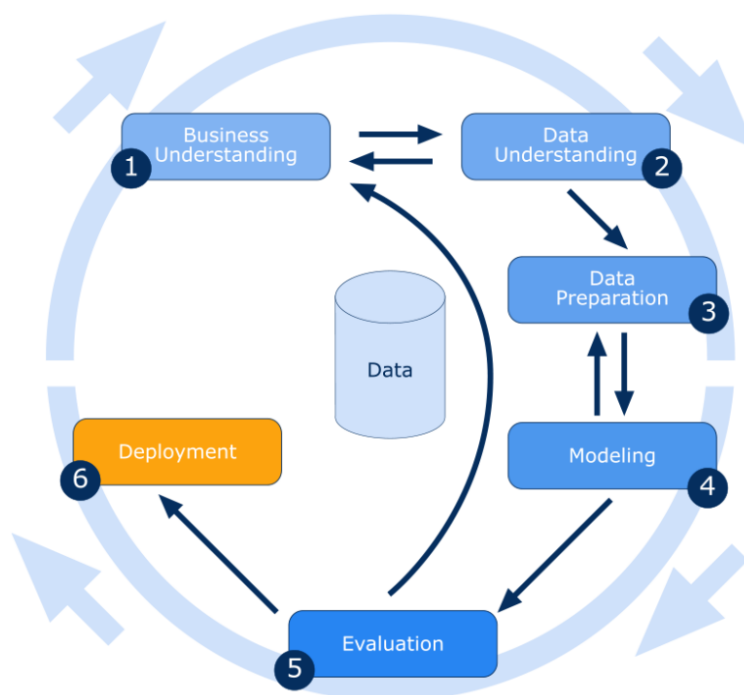


Imagem 1: Fluxograma do CRISP DM.

1. **Entendimento do negócio:** Durante essa etapa, é importante identificar as necessidades e expectativas, principalmente aqueles que irão lidar com os dados e informações obtidos para posterior análise e tomada de decisão. No projeto, dividimos a equipe e buscamos, em primeiro plano, entender e estudar a área que o cliente tem a demanda, nesse caso, a área da saúde.
2. **Entendimento dos dados:** É a parte onde exploramos e tratamos os dados. Entendendo os dados disponíveis e dados vazios na nossa base, fazendo avaliação dos dados disponíveis e da qualidade dos mesmos. Verificando se a volumetria dos dados atende ao estudo do negócio, referindo-se a quantidade que será utilizada, juntamente com o levantamento de hipóteses consequentes da exploração dos dados.
3. **Preparação dos dados:** A etapa em que os dados serão tratados, como dados em formatos de “string” deixam de ser categóricos, ou seja, são tratados para o funcionamento dos modelos de predição, chamados de normalização, padronização.

Como mostrado no fluxo apresentado acima, é uma etapa que se repete até que os resultados do modelo sejam satisfatórios.

4. **Modelagem:** Essa parte acontece após o tratamento de dados e se inicia o processo de mineração dos dados, onde esses dados passam por processos com o objetivo de serem encontrados padrões que fazem um determinado evento acontecer, escolhendo técnicas mais adequadas de modelagem com base na mineração. Quando iniciamos essa etapa, notamos a necessidade de retornar a etapa anterior para que os dados sejam tratados novamente e refinados, como mostrado no fluxo dos processos.
5. **Avaliação:** Nessa fase, o processo da fase anterior se resulta em modelos de predição. Nesse momento, é importante a observação de algumas métricas de avaliação, como por exemplo a acurácia e a precisão, escolhemos os modelos com melhor desempenho baseado nos dados já tratados e validados. Com a observação de critérios não atendidos, voltamos a primeira etapa de entendimento do negócio, até que o modelo atenda o critério de sucesso com uma solução concreta.
6. **Implementação:** Com o modelo resultante atendendo a demanda e necessidade do cliente, ele é implementado pelo parceiro e apresentando possíveis aderências, buscando ser interpretável.

Abaixo encontram-se listadas as **principais ferramentas** utilizadas pelo grupo para realizar o projeto, seguidas de especificação de seus respectivos usos dentro desse mesmo contexto.

- Google Collaboratory (Colab) - desenvolvimento do código (escrito na linguagem Python).
- GitHub - setor de entregas do código e da documentação.
- Google Docs - documentação.
- Google Drive - repositório da base de dados.

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 2: 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 3: 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 desenvolvimento	Modelo não conseguir contemplar todas as expectativas dos integrantes	Falta de dados suficientes	Integrantes não se adaptarem com o uso da ferramenta	Pacientes conseguem ser tratados de uma maneira mais adequada	Redução do tempo de escolha de um tratamento	Modelo conseguir contemplar todas as expectativas do cliente		+90%	
	75%		Falta de alinhamento dos integrantes sobre o tema	Médicos não se adaptarem com o uso da ferramenta	Prazo final curto demais	Criação de emprego na manutenção da solução produzida	Aumentar o networking entre os membros do grupo			75%	
	50%		Bugs do Colab	Não entendimento de tópicos essenciais para a confecção do modelo	Modelo com falta de precisão	Falta de comunicação entre os integrantes	Aumentar a consciência sobre câncer de mama	Diminuição de custos		50%	
	25%		Mai funcionamento dos computadores dos integrantes	Falta de apoio do corpo docente	Parceiro não gostar do produto final		Tratamento mais personalizado			25%	
	10%									10%	
Impacto											

Imagem 4: 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	Personalidade	Dores	Necessidades
Idade	42 anos	<ul style="list-style-type: none"> • Extrovertida • Intuitiva • Emocional • Criativa 	<ul style="list-style-type: none"> • Indicação de tratamentos ineficientes. • Tratamentos não personalizados. • Gasto de tempo, atrapalhando a indicação de tratamentos. 	<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.
Ocupação	Oncologista			
Localização	São Paulo, SP			
Educação	Mestrado			

Imagem 5 - 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 6 - 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 7 - 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_seguimento
idade_diagnostico
gestacao_idade
idade_primeira_menstruacao
tempo_amamentacao

Categórico:

escolaridade
sexo
raca
ultima informação 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

```

b) Estatística descritiva das colunas.

A fim de tornar a análise de dados mais fácil e eficiente, optamos por realizar uma análise descritiva em uma tabela de cada vez, explorando ao máximo suas colunas. Para visualizar as tabelas dentro da ferramenta Colab, utilizamos a função 'read' da biblioteca pandas.

```

df_demograficos = pd.read_csv('/content/drive/MyDrive/Colab - DEV/BDIPMamaV11-INTELI/DemograficosTt_DATA_LABELS_2023-01-24_1922.csv')
df_registro_tumo = pd.read_csv('/content/drive/MyDrive/Colab - DEV/BDIPMamaV11-INTELI/RegistroDeTumo_DATA_LABELS_2023-01-24_1924.csv')
df_histopatologia = pd.read_csv('/content/drive/MyDrive/Colab - DEV/BDIPMamaV11-INTELI/Histopatologia_DATA_LABELS_2023-01-24_1924.csv')
df_pesoEaltura = pd.read_csv('/content/drive/MyDrive/Colab - DEV/BDIPMamaV11-INTELI/PesoEAltura_DATA_LABELS_2023-01-24_1926.csv')

```

Imagem 8: Função read aplicada em todas as tabelas.

Nesse primeiro momento, utilizamos a função *shape* para identificar a quantidade de linhas e colunas, respectivamente, que existem naquele *dataframe*. Além do *shape*, utilizamos a função *info()* para observar todos os nomes das colunas e seus tipos (*int*, *float* ou *object*). Também aplicamos a função *describe()* que tem o objetivo de mostrar as estatísticas descritivas, como a moda, mediana, percentil, desvio padrão e a média de cada coluna.

1º Passo:

```

df_demograficos.shape

(4272, 24)

```

Imagem 9: Utilização da função shape na tabela df_demograficos.

2º Passo:

```
df_demograficos.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 4272 entries, 0 to 4271
Data columns (total 24 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   record_id                                4272 non-null   int64
1   escolaridade                             3883 non-null   float64
2   idade_diagnostico                        4092 non-null   float64
3   sexo                                     4125 non-null   float64
4   raca                                     221 non-null    float64
5   ultima_informacao_paciente              4270 non-null   float64
6   tempo_seguimento                        4270 non-null   float64
7   ja_gravida                              1013 non-null   float64
8   gestacao_idade                          897 non-null    float64
9   tempo_amamentacao                       688 non-null    float64
10  grau_parentesco_mama_primeiro_1_vez     4272 non-null   int8
11  grau_parentesco_mama_primeiro_mais_vezes 4272 non-null   int8
12  grau_parentesco_mama_segundo_1_vez      4272 non-null   int8
13  grau_parentesco_mama_segundo_mais_vezes 4272 non-null   int8
14  idade_primeira_menstruacao              1025 non-null   float64
15  uso_anticoncepcional                    4272 non-null   int8
16  atividade_fisica                         305 non-null    float64
17  consumo_alcool                           204 non-null    float64
18  historico_cancer                        190 non-null    float64
19  grau_parentesco_primeiro                 4272 non-null   int8
20  grau_parentesco_segundo                  4272 non-null   int8
21  grau_parentesco_terceiro                  4272 non-null   int8
22  tratamento                              2863 non-null   float64
23  anti_her2_neoadjuvante                   1134 non-null   float64
dtypes: float64(15), int64(1), int8(8)
memory usage: 600.8 KB
```

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

3º Passo:

```
df_demograficos.describe()

      record_id  escolaridade  idade_diagnostico  sexo  raca  ultima_informacao_paciente  tempo_seguimento
count  4272.000000    3883.000000    4092.000000  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      98.000000    1.000000    3.000000      3.000000    4503.000000
```

Imagem 11: 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.

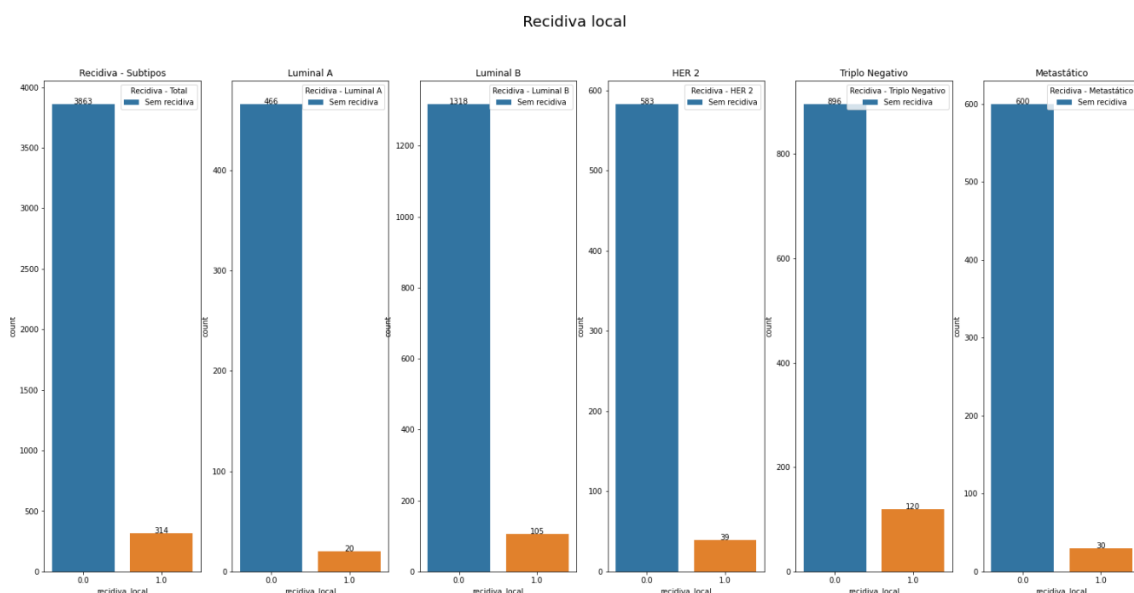


Imagem 12: Representação gráfica da Recidiva Local.

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 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.

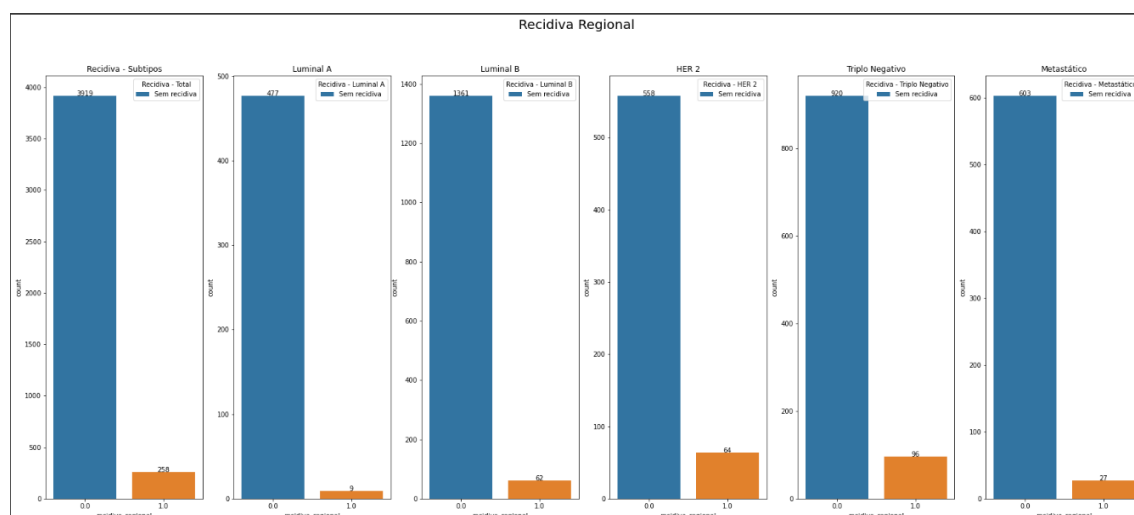


Imagem 13: Representação gráfica da Recidiva Regional.

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. Nesse caso,

dos 4177 pacientes, 744 apresentaram recidivas à distância.

Dentre os subtipos de tumor, o que mais apresenta esse tipo de recidiva é o triplo negativo com cerca de 23% de recorrência. Ou seja, a cada 100 pacientes diagnosticados com este subtipo 23 apresentam recidiva em um local distante do tumor inicial.

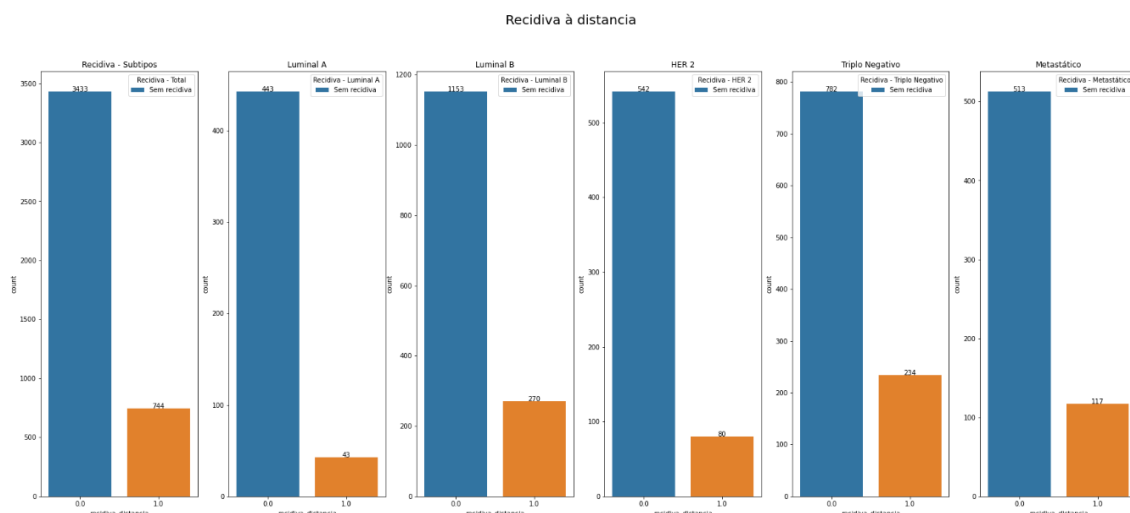


Imagem 14: Representação gráfica da Recidiva à distância.

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 15: 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 16: Código desenvolvido pela equipe.



Imagem 17: 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 18: 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.

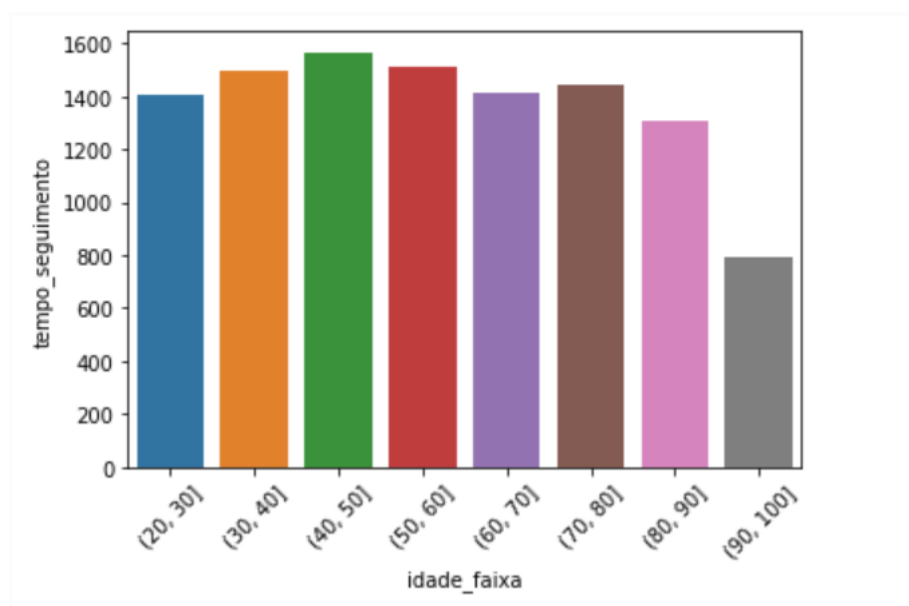


Imagem 19: Gráfico de tempo de seguimento em relação à idade. Elaborado pela equipe.

4.3. Preparação dos Dados e Modelagem

1. Modelo supervisionado:

a) Modelagem para o problema (proposta de features com a explicação completa da linha de raciocínio).

Utilizamos o algoritmo de aprendizagem de máquina não supervisionado *K-Means* para agrupar os tratamentos com resultados semelhantes em clusters específicos, a fim de criar o modelo. Essa abordagem substituiu a necessidade de usar vários "if's" para classificar entre tratamentos bons e ruins. A partir dos dois *clusters* identificados, treinamos um modelo supervisionado utilizando o *dataframe* dos tratamentos que tiveram resposta positiva. Com base nesses exemplos de tratamentos "bons", o modelo supervisionado pode prever qual tratamento será mais eficaz para futuros pacientes.

```
from sklearn.cluster import KMeans

km = KMeans( n_clusters = 2, init = 'random', max_iter = 300, n_init = 100, random_state = 52 )
```

Imagem 20: Importação do algoritmo *K-Means* da biblioteca *scikit-learn*.

Ao modelar o problema, selecionamos as variáveis que melhor explicam a variabilidade dos dados e são relevantes para a classificação dos tratamentos como "bons". Algumas variáveis potenciais incluem 'idade_diagnostico', 'primeiro_IMC', 'última_informação_paciente', 'tempo_seguimento', 'recidiva_distancia' e 'recidiva_regional'. Essas variáveis podem ser úteis para entender como diferentes características do paciente ou do tratamento podem influenciar na eficácia do tratamento, ajudando a identificar os melhores tratamentos para futuros pacientes.

```
df[['idade_diagnostico', 'primeiro_IMC', 'ultima_informacao_paciente', 'tempo_seguimento', 'recidiva_distancia', 'recidiva_regional', 'recidiva_local']]
```

Imagem 21: Variáveis úteis para a escolha do tratamento.

b) Métricas relacionadas ao modelo (conjunto de testes, pelo menos 3).

A matriz de confusão é uma tabela que apresenta as frequências de cada classe em um modelo. Ela contém quatro valores: *True Positive* (TP), *False Negative* (FN), *False Positive* (FP) e *True Negative* (TN).

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Imagem 22: Tabela que mostra a matriz de confusão em português.

True Positive (TP) representa os casos em que o modelo previu corretamente a classe. Por exemplo, quando um modelo previu que um paciente tinha câncer e ele realmente tinha câncer.

False Positive (FP) representa os casos em que o modelo previu incorretamente a classe. Por exemplo, quando um modelo previu que um paciente tinha câncer, mas na verdade ele não tinha.

False Negative (FN) representa os casos em que o modelo previu incorretamente a classe que não estamos buscando. Por exemplo, quando um modelo previu que um paciente não tinha câncer, mas na verdade ele tinha.

True Negative (TN) representa os casos em que o modelo previu corretamente a classe que não estamos buscando. Por exemplo, quando um modelo previu que um paciente não tinha câncer e ele realmente não tinha.

A partir desses valores, é possível calcular outras métricas, como acurácia, precisão e recall.

A acurácia representa a quantidade de acertos em relação a todas as predições e é dada pela soma da diagonal maior da matriz de confusão dividida pela quantidade de predições realizadas (corretamente ou incorretamente).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Imagem 23: Fórmula da acurácia.

A precisão determina a proporção entre a quantidade de predições verdadeiras positivas e as predições que o modelo previu como positivas.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Imagem 24: Fórmula da precisão.

O *recall* é responsável por determinar a proporção entre a quantidade de predições que o modelo previu corretamente e os casos que realmente eram positivos.

$$\text{Recall} = \frac{\text{True Positive}(TP)}{\text{True Positive}(TP) + \text{False Negative}(FN)}$$

Imagem 25: Fórmula do *Recall*.

O *F1-Score* é uma métrica que representa uma média harmônica entre a precisão e o *recall*.

$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

Imagem 26: Fórmula do *F1-Score*.

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica que ajuda a entender a qualidade do modelo entre a taxa de verdadeiros positivos e a taxa de falsos positivos.

O AUC (*Area Under the Curve*) é uma métrica utilizada para quantificar a qualidade da curva ROC, variando entre 0 e 1. Um valor de 0 indica um mau desempenho na diferenciação entre as classes positivas e negativas, enquanto um valor de 1 indica um bom desempenho.

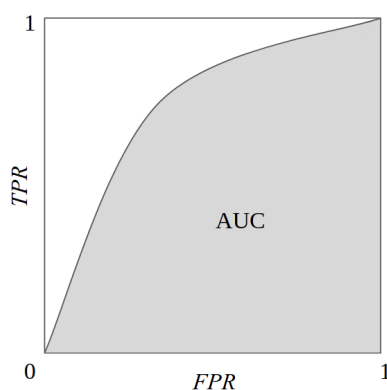


Imagem 27: Gráfico de AUC.

Com base nesses dados, foi possível treinar um modelo supervisionado, que utiliza algoritmos de aprendizagem de máquina para aprender a identificar padrões e relações entre as variáveis de entrada e a variável de saída. O modelo foi avaliado usando métricas como acurácia, precisão, *recall* e *F1-score*. Além disso, o modelo foi testado em um conjunto de dados separado para verificar se ele é capaz de generalizar para novos dados e evitar o *overfitting* (quando a acurácia do modelo é alta no conjunto de treinamento, mas baixa no conjunto de testes).

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken	Precision	Recalls
AdaBoostClassifier	0.68	0.67	0.67	0.68	0.59	0.66	0.63
ExtraTreesClassifier	0.67	0.67	0.67	0.67	0.50	0.66	0.62
RandomForestClassifier	0.67	0.66	0.66	0.67	1.70	0.65	0.60
RidgeClassifier	0.66	0.66	0.66	0.66	0.04	0.66	0.60
LinearSVC	0.66	0.66	0.66	0.66	0.52	0.47	0.00
RidgeClassifierCV	0.66	0.66	0.66	0.66	0.10	0.59	0.46
LinearDiscriminantAnalysis	0.66	0.66	0.66	0.66	0.16	0.00	0.00
LogisticRegression	0.66	0.66	0.66	0.66	0.09	0.65	0.62
CalibratedClassifierCV	0.66	0.66	0.66	0.66	1.36	0.65	0.59
SVC	0.66	0.66	0.66	0.66	0.81	0.65	0.60
LGBMClassifier	0.66	0.65	0.65	0.66	7.67	0.52	0.38
XGBClassifier	0.65	0.65	0.65	0.65	4.95	0.62	0.65
NearestCentroid	0.64	0.64	0.64	0.64	0.07	0.46	0.59
BaggingClassifier	0.64	0.64	0.64	0.64	0.37	0.47	0.99
BernoulliNB	0.64	0.64	0.64	0.64	0.02	0.64	0.52
NuSVC	0.62	0.62	0.62	0.62	0.80	0.55	0.47
DecisionTreeClassifier	0.61	0.61	0.61	0.61	0.08	0.47	0.00
KNeighborsClassifier	0.60	0.60	0.60	0.60	0.07	0.47	0.96
SGDClassifier	0.60	0.60	0.60	0.60	0.15	0.39	0.28
Perceptron	0.58	0.58	0.58	0.58	0.08	0.58	0.59
LabelSpreading	0.57	0.56	0.56	0.56	0.32	0.46	0.46
LabelPropagation	0.57	0.56	0.56	0.56	0.28	0.00	0.00
PassiveAggressiveClassifier	0.55	0.55	0.55	0.54	0.03	0.00	0.00
ExtraTreeClassifier	0.54	0.54	0.54	0.54	0.02	0.54	0.59
GaussianNB	0.47	0.50	0.50	0.31	0.03	0.57	0.72
DummyClassifier	0.53	0.50	0.50	0.37	0.02	0.00	0.00
QuadraticDiscriminantAnalysis	0.48	0.50	0.50	0.44	0.15	0.46	0.94

Imagem 28: Tabela de métricas e modelos.

c) Apresentar o primeiro modelo candidato, e uma discussão sobre os resultados deste modelo (discussão sobre as métricas para esse modelo candidato).

Com o modelo treinado e avaliado, fazer previsões sobre quais tratamentos (neoadjuvante ou adjuvante) serão mais eficazes para novos pacientes, com base em suas características e histórico médico. Isso significa que o modelo pode sugerir ao médico quais tratamentos apresentam maiores chances de sucesso para cada paciente, aumentando a precisão do diagnóstico e melhorando a qualidade do atendimento e dos resultados obtidos.

O nosso primeiro modelo candidato é o 'AdaBoostClassifier'. A acurácia do modelo foi de 0.68, o que significa que ele acertou 68% das previsões. A precisão foi de 0.66, o *recall* foi de 0.63 e o *F1-score* foi de 0.68. Esses resultados indicam que o modelo tem um bom desempenho na previsão de quais tratamentos são mais eficazes, com uma boa relação entre acurácia e *F1-Score*.

4.4. Comparação de Modelos

a) Escolha da métrica e justificativa.

Ao escolhermos a métrica de avaliação para nosso modelo, consideramos que a acurácia é a mais importante. Isso ocorre porque ambos os tratamentos não possuem uma hierarquia de valor e, portanto, as predições de verdadeiro positivo e verdadeiro negativo têm a mesma relevância. A acurácia nos ajuda a avaliar quão bem o modelo está selecionando o tratamento mais adequado para cada paciente. No entanto, reconhecemos que outras métricas também são importantes para garantir que o modelo não esteja enviesado e para avaliar sua qualidade geral.

b) Modelos otimizados.

Nós escolhemos três modelos para avaliação, sendo eles o *AdaBoost Classifier*, *Random Forest Classifier* e *Light Gradient Boosting Machine* (LGBM). Optamos por otimizar esses modelos utilizando algoritmos de otimização de hiperparâmetros, como *Grid Search* e *Random Search*. O *AdaBoost Classifier* utiliza uma técnica de boosting para selecionar o modelo preditivo mais adequado de acordo com o padrão de informações. Já o *Random Forest Classifier* seleciona os melhores recursos e cria múltiplas árvores de decisão, cada uma selecionando combinações aleatórias das características escolhidas e, em seguida, seleciona a previsão mais comum entre as árvores. No LGBM, dividimos os dados em subconjuntos menores e para cada conjunto é instanciada uma árvore de decisão. Além disso, utilizamos o método de 'gradient boosting' para melhorar o desempenho de cada árvore. O resultado final foi obtido por meio da média ponderada das predições de cada árvore.

c) Definição do modelo escolhido e justificativa.

Durante o ajuste de hiperparâmetros, o algoritmo de otimização *Grid Search* foi priorizado devido ao seu melhor desempenho em comparação com outras opções testadas.

RANDOM FOREST CLASSIFIER

Os hiperparâmetros previamente testados para o modelo *Random Forest Classifier*, que apresentou melhor desempenho, são listados a seguir:

```
params = {  
    'n_estimators': [50, 100, 150],  
    'max_depth': [5, 10, 15],  
    'max_features': ['sqrt', 'log2', None],  
    'random_state': [73]  
}
```

Imagem 29: Hiperparâmetros do *Random Forest Classifier*.

ADA BOOST CLASSIFIER

Os hiperparâmetros previamente testados para o modelo *AdaBoost Classifier*, que obteve a segunda melhor performance, são listados a seguir:

```
params = {  
    'n_estimators': [50, 100, 150],  
    'learning_rate': [0.01, 0.1],  
    'algorithm': ['SAMME', 'SAMME.R'],  
    'random_state': [73]  
}
```

Imagem 30 : Hiperparâmetros do *AdaBoost Classifier*.

LIGHT GRADIENT BOOSTING MACHINE

Os hiperparâmetros previamente testados para o modelo *AdaBoost Classifier*, que obteve a segunda melhor performance, são listados a seguir:

```
params = {  
    'max_depth': [5, 10],  
    'learning_rate': [0.01, 0.05],  
    'n_estimators': [50, 100],  
    'random_state': [73]  
}
```

Imagem 31 : Hiperparâmetros do *Light Gradient Boosting Machine*.

MÉTRICAS APÓS A OPTIMIZAÇÃO DOS HIPERPARÂMETROS

Nós listamos abaixo as acurácias apresentadas pelos modelos após a escolha dos hiperparâmetros mais adequados, uma vez que consideramos essa métrica como sendo de maior importância para a validação do modelo. Além disso, destacamos que a escolha do *Grid Search* em detrimento do *Random Search* também é comprovada pela comparação das acurácias de cada método otimizador.

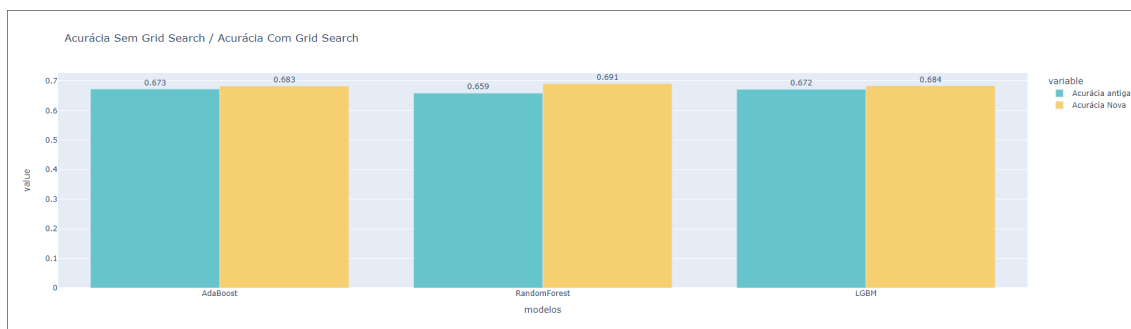


Imagem 32 : Gráfico representa as acurácias sem e com o *Grid Search*.

Modelo utilizado	Acurácia antes do <i>Grid Search</i>	Acurácia depois do <i>Grid Search</i>
<i>AdaBoost Classifier</i>	0.673	0.683
<i>Random Forest Classifier</i>	0.659	0.691
<i>LGBM</i>	0.672	0.684

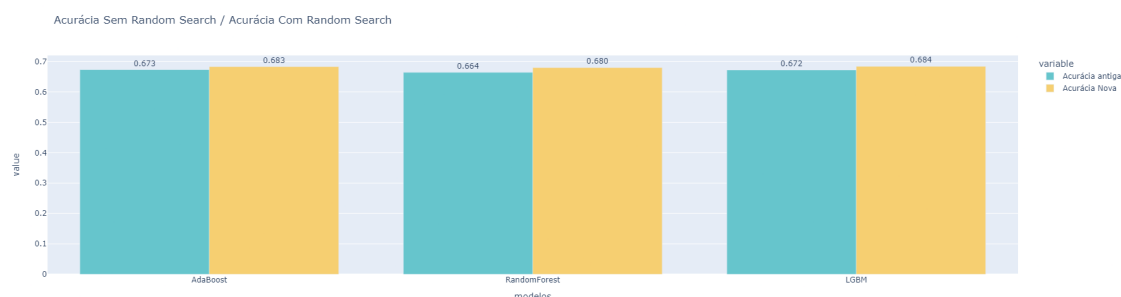


Imagem 33 : Gráfico representa as acurácias sem e com *Random Search*.

Modelo utilizado	Acurácia antes do <i>Random Search</i>	Acurácia depois do <i>Random Search</i>
<i>AdaBoost Classifier</i>	0.673	0.683
<i>Random Forest Classifier</i>	0.664	0.680
<i>LGBM</i>	0.672	0.684

Concluimos que o modelo escolhido foi o *Random Forest Classifier* com 69,1% de acurácia, utilizando o *Grid Search* como método de otimização de hiperparâmetros. Esse modelo apresentou a melhor performance em relação à métrica de acurácia, justificando sua escolha.

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 predições.

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.

"CRISP DM - O que é e como usar?", 2020. Disponível em:

< <https://www.linkedin.com/pulse/crisp-dm-o-que-%C3%A9-e-como-usar-rodrigo-ribeiro/?originalSubdomain=pt> > Acesso em 08/03/2023.

Anexos