



## Controle do Documento

### Histórico de revisões

Data	Autor	Versão	Resumo da atividade
01/02/2023	Yago Phellipe	0.1	Preenchimento da seção 4.1.3.
02/02/2023	José Vitor Marcelino	0.2	Implementação de Personas 4.1.6
06/02/2023	José Vitor Marcelino	0.3	Rascunho da Introdução 1.0 (Pré-Validação com a professora)
06/02/2023	Emely Vitória	0.4	Preenchimento da seção 4.1.1
07/02/2023	Marcos Teixeira	0.5	Preenchimento da seção 4.1.5
08/02/2023	Marcos Teixeira	0.6	Preenchimento da seção 4.1.5
08/02/2023	Yuri Toledo	0.7	Preenchimento da seção 4.1.7
09/02/2023	Marcos Teixeira	0.8	Preenchimento da seção 4.1.3
10/02/2023	Emely Vitória	0.9	Preenchimento da seção 2.1
20/02/2023	Yuri Toledo	1.1	Correção dos erros apontados pela professora na última sprint
23/02/2023	Vivian Shibata	1.2	Preenchimento da LGPD, análise SWOT e ABNT
26/02/2023	Vivian Shibata	1.3	Preenchimento das hipóteses na compreensão dos dados
26/02/2023	Yago Phellipe	1.4	Preenchimento da questão 1 letra A e B da seção 4.2
26/02/2023	Emely Vitória, Yuri Toledo	1.5	Preenchimento da seção 4.2.1.2 letra A
27/02/2023	Daniel Dávila	1.6	Reescrita dos itens 1 e 2; preenchimento do item 3

07/03/2023	José Vitor Marcelino, Vivian Shibata	1.7	Reescrita do item 3 e preenchimento da seção 4.3 letra A
09/03/2023	Yuri Toledo	1.8	Preenchimento da letra b da seção 4.3
11/03/2023	Yuri Toledo	1.9	Preenchimento da letra c da seção 4.3

# Sumário

<b>1. Introdução</b>	<b>4</b>
<b>2. Objetivos e Justificativa</b>	<b>5</b>
2.1. Objetivos	5
2.2. Proposta de Solução	5
2.3. Justificativa	5
<b>3. Metodologia</b>	<b>6</b>
<b>4. Desenvolvimento e Resultados</b>	<b>7</b>
4.1. Compreensão do Problema	7
4.1.1. Contexto da indústria	7
4.1.2. Cinco forças de Porter	7
4.1.3. Análise SWOT	7
4.1.4. Planejamento Geral da Solução	7
4.1.5. Value Proposition Canvas	7
4.1.6. Matriz de Riscos	7
4.1.7. Personas	8
4.1.8. Jornadas do Usuário	8
4.1.9 Política de privacidade para o projeto de acordo com a LGPD	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
<b>5. Conclusões e Recomendações</b>	<b>13</b>

<b>6. Referências</b>	<b>14</b>
-----------------------	-----------

<b>Anexos</b>	<b>15</b>
---------------	-----------

# 1. Introdução

O Instituto do Câncer de São Paulo (ICESP) é um dos maiores hospitais especializados no tratamento de câncer na América Latina.

Fruto de uma parceria entre a Faculdade de Medicina da USP (FMUSP) e o Governo de São Paulo, o Instituto foi inicialmente idealizado em 1987 e oficialmente fundado em 2008. O ICESSP é inteiramente dedicado ao paciente oncológico, principalmente naquilo que concerne o tratamento do câncer, mas também naquilo que concerne desenvolver e aperfeiçoar medicamentos e tratamentos por meio de pesquisas científicas.

Por virtude do fato de que o ICESSP é uma instituição pública sem fins lucrativos - isto é, devotada puramente ao aumento do bem-estar social - inexistem possíveis concorrentes: a assistência médica que providencia é gratuita\*, e as pesquisas científicas que produz pertencem ao domínio público. Notavelmente, cada vez mais, instituições como o ICESSP buscam investir em tecnologia aperfeiçoadora da eficiência e efetividade imanente aos processos hospitalares.

O mais recente desses investimentos é o desenvolver de inteligência artificial que, com relativo alto grau de acurácia, prevê o tratamento probabilisticamente ideal para cada paciente de câncer de mama - uma probabilidade que é adquirida por meio da meticulosa análise de múltiplos dados subproduto de detalhados relatórios sobre milhares de pacientes.

Este documento objetiva registrar tal desenvolvimento.

\* (porém ocasionalmente o Hospital presta serviços médicos pagos para complementar custos operacionais)

## 2. Objetivos e Justificativa

### 2.1. Objetivos

O principal objetivo do parceiro de negócio (ICESP) consiste em aumentar a eficiência e a efetividade da decisão do médico sobre qual tratamento deve ser aplicado para cada paciente de câncer de mama.

### 2.2. Proposta de Solução

Para o alcance de tal objetivo, será desenvolvido modelo preditivo que auxilie o médico na tomada de tal decisão, e para isso serão realizadas, em síntese, as seguintes tarefas:

1. **Analisar** uma ampla variedade de dados, essencialmente divisíveis em "clínicos" (informações sobre saúde, histórico médico e medicações) e "demográficos" (informações sobre idade, gênero, renda, educação, et cetera).
2. Com base em tal análise, **filtrar** as informações essenciais do paciente de maneira que seja prevista com relativo alto grau de acurácia as chances de saída futura do paciente.
3. Com base em tal filtragem, **classificar** dados de pacientes prévios visando identificar qual a melhor forma\* de realizar o tratamento de câncer de mama.
4. Com base em tal classificação, **detectar** padrões capazes de indicar aos médicos o tratamento probabilisticamente ideal para cada paciente.
5. **Realizar** tal indicação.

\*A melhor forma pode ser a adjuvante (1.º cirurgia e 2.º terapia), ou a neoadjuvante (1.º quimioterapia e 2.º cirurgia).

### 2.3. Justificativa

A maior dificuldade enfrentada por médicos de câncer de mama consiste em decidir qual é o tratamento ideal para cada paciente, pois ele varia significativamente entre cada um. Desprovidos de uma métrica acurada para medir tal variância, os médicos são forçados a recorrer à análise manual dos dados do paciente para decidir qual é o tratamento mais adequado. Posto que inerente à realização manual de tarefas é a drástica redução de eficiência e de efetividade, pode-se concluir que é necessária uma solução.

Tal solução é o desenvolvimento de um modelo preditivo que indica qual tratamento possui maior probabilidade de sucesso - uma probabilidade que é adquirida por meio da meticulosa análise de múltiplos dados subproduto de detalhados relatórios sobre milhares de pacientes.

### 3. Metodologia

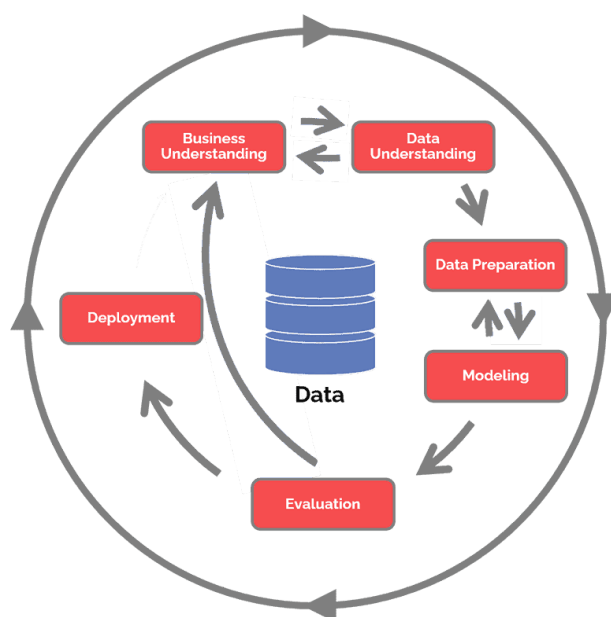


Figura 1 - Metodologia CRISP-DM

O CRISP-DM (Cross Industry Standard Process for Mining Data) é uma metodologia que consiste em um conjunto de práticas para mineração de dados. Tais práticas resumem-se no seguir de 6 passos, sendo eles:

- 1.º Passo - Entendimento do Negócio: consiste em compreender o objetivo do projeto e o problema a ser resolvido, sempre evitando vieses inconscientes. É importante captar todos os detalhes - não apenas aqueles relativos ao problema e ao projeto - mas também aqueles relativos à empresa, posto que entender a estratégia é essencial.
- 2.º Passo - Entendimento dos Dados: os dados são dissecados, organizados, analisados, filtrados, documentados, estudados, e profundamente compreendidos; "Verbrennen musst du dich wollen in deiner eigenen Flamme. Wie wolltest du neu werden, wenn du nicht zuvor Asche geworden bist!"<sup>1</sup>.
- 3.º Passo - Preparação dos Dados: é o passo que antecede a construção dos modelos. É a fase mais complexa da CRISP-DM, demandando aproximadamente de 70 a 90% do tempo total do projeto, e é focada no pré-processamento de dados. Essencialmente, tal



pré-processamento significa tratar os dados conforme interesse e necessidade, excluir dados anômalos e vazios, normalizar, e padronizar.

- 4.º Passo - Modelagem: etapa em que é escolhido o tipo de modelagem ideal para a base de dados. Tal tarefa é realizada por meio de ferramentas computacionais, objetivando resolver o problema analisado no 1.º Passo - Entendimento do Negócio.
- 5.º Passo - Avaliação do Modelo: análise do resultado do modelo, verificação se tal resultado corresponde a expectativas pré-estabelecidas pela equipe. Em caso negativo, é recomenda-se reavaliar as etapas anteriores, visando o aprimoramento do modelo preditivo.
- 6.º Passo - Deployment: consiste em colocar o modelo em ação de maneira que valor seja agregado para o negócio. O modelo costuma ficar armazenado na nuvem ou em servidores locais do cliente.

Além disso, é importante mencionar que, como demonstra a Figura 1, possíveis falhas durante o processo pressupõem a revisão de um ou mais passos realizados previamente. Tal prática assegura a constante aprimoração do processo de análise exploratória dos dados.

<sup>1</sup>Nietzsche

## 4. Desenvolvimento e Resultados

### 4.1. Compreensão do Problema

#### 4.1.1. Contexto da indústria

Após meticulosa examinação da análise de mercado, mostrou-se impossível identificar qualquer tipo de concorrente direto ao parceiro de negócios (ICESP). Estima-se que tal fato seja subproduto do posicionamento do parceiro como instituição pública sem fins lucrativos devotada puramente ao servir da sociedade: a assistência médica é gratuita\* e as pesquisas científicas pertencem ao domínio público.

Apesar disso, é necessário mencionar a existência de outras empresas que atuam na mesma área, como o Hospital A.C. Camargo, especializado no tratamento e pesquisa do câncer (CA); O Centro oncológico Família Dayan - Daycoval, pertencente ao Hospital Israelita Albert Einstein; e o INCA, um órgão vinculado ao Ministério da Saúde responsável pela prevenção e controle do câncer no Brasil.

Consequente à natureza funcional do ICESP é a incessante busca pela implementação de novas soluções tecnológicas. Exemplos disso são: a procura por equipamentos mais avançados, a implementação de tratamentos mais eficazes e menos prejudiciais, as colaborações com iniciativas privadas, e a crescente internacionalização que objetiva tornar-lhe um centro educacional internacionalmente reverenciado<sup>1</sup>; sem uma sombra de dúvidas sequer, determinações que destacam o Hospital no ambiente de mercado ao qual pertence.

\*(porém ocasionalmente o Hospital presta serviços médicos pagos para complementar custos operacionais)

<sup>1</sup>(USP, 2021)

#### 4.1.2. Cinco forças de Porter

Rivalidade entre os concorrentes: inexistente, pois o ICESP é uma instituição pública sem fins lucrativos, apesar de que existem empresas que atuam na mesma área. (fato apresentado em mais detalhes no primeiro parágrafo do item 4.1.1)

Poder de barganha de clientes: variável de acordo com o nível socioeconômico e da disponibilidade de opções para cada cliente. Por outro lado, a instituição possui grande demanda, o que acaba por limitar o poder de negociação dos clientes.

Poder de barganha de fornecedores: quanto a fornecedores de equipamento, o poder de barganha é alto, pois o ICESP necessita de produtos tecnologicamente vanguardistas e monetariamente custosos; certas medicações e determinados equipamentos não são amplamente disponíveis, com produção e preço amplamente controlados por pequeno grupo de empresas. Quanto a fornecedores de dados (outros hospitais), o poder de barganha também é alto, pois o fornecimento pode ser recusado.

Ameaças de produtos substitutos: baixa, pois não há preocupação com a entrada de novos competidores, pois as instituições na mesma área de atuação do Hospital das Clínicas não são vistas como concorrentes, mas sim como parceiros. A FMUSP é reconhecida como uma referência na área da saúde, e qualquer nova empresa que surja no mercado não seria capaz de competir com ela a prazo estimável.

Ameaças de novos entrantes: novamente, é baixa, pois, reiterando, não há preocupação com a entrada de novos competidores, pois as instituições na mesma área de atuação do Hospital das Clínicas não são vistas como concorrentes, mas sim como parceiros. A FMUSP é reconhecida como uma referência na área da saúde, e qualquer nova empresa que surja no mercado não seria capaz de competir com ela a prazo estimável.

### 4.1.3. Análise SWOT

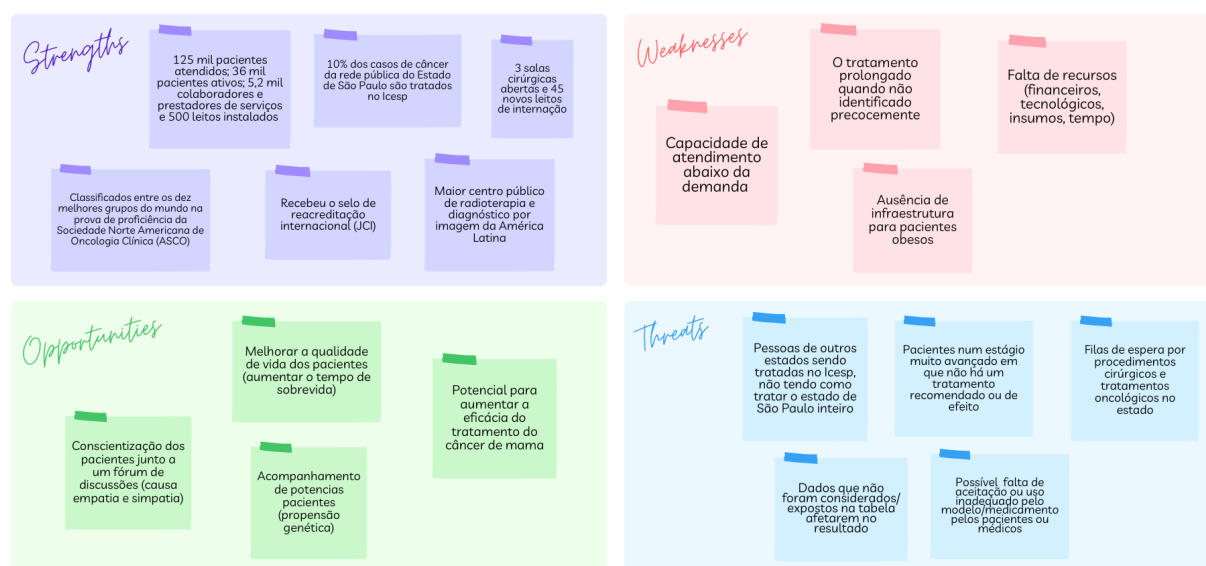


Figura 2 - Matriz SWOT

Fonte: Elaboração própria

#### Forças:

De acordo com sua plataforma oficial<sup>1</sup>, o Instituto do Câncer do Estado de São Paulo é responsável por tratar cerca de 10% dos casos de câncer da rede pública do Estado de São Paulo. Contando com 5,2 mil colaboradores e prestadores de serviços, o Instituto atendeu um total de 121 mil pacientes desde a sua inauguração em maio de 2008. Em sua capacidade máxima disponibiliza 490 leitos, 85 dos quais são da Unidade de Terapia Intensiva (UTI).

Após ampliações infraestruturais recentes, foram abertas 3 novas salas cirúrgicas e 45 novos leitos de internação, sendo 15 da UTI. Além de possibilitar o atendimento de 1250 novos pacientes, também permite a realização de 840 cirurgias adicionais - equivalente a um incremento de 20% do previsto para o período. Essa iniciativa visa reduzir, dentro dos primeiros meses da ação, a fila de pacientes oncológicos no Estado de São Paulo em 40%<sup>2</sup>.

Em 2014, ICESP tornou-se o primeiro hospital da rede pública da capital a receber o selo da Joint Commission International (JCI) - uma certificação internacional que reconhece excelência em atendimentos e serviços oferecidos à população. De três em três anos, o selo passa por um processo de verificação. A mais recente ocorreu em 2020<sup>3</sup>.

Quanto ao ensino que oferece, o ICESP também se destaca: no Programa de Residência Médica em Cancerologia Clínica, originado em 1998 como primeiro do Brasil, formou mais de 170 médicos oncologistas. Reconhecido tanto nacionalmente quanto internacionalmente, é, em sua categoria, um dos maiores programas do país.

Nos últimos anos, os formandos foram sistematicamente classificados entre os dez melhores grupos do mundo na prova de proficiência da Sociedade Norte Americana de Oncologia Clínica (ASCO). No ano de 2022, os residentes do segundo e terceiro ano alcançaram o melhor desempenho no exame anual da mesma instituição. Fora de todos os competidores, o grupo obteve a maior média, e atingiu sua maior pontuação em relação a avaliações de anos anteriores. Tais resultados posicionam os profissionais do ICESP dentre os melhores do mundo.

<sup>1</sup>(ICESP, 2022)

<sup>2</sup>(ICESP, 2023)

<sup>3</sup>(GOVERNO, 2021)

<sup>4</sup>(ICESP, 2023)

### **Fraquezas:**

A maior fraqueza do ICESP é a falta de recursos; a principal causa tratamentos falhos e de demanda insuficientemente atendida. Tratamentos de imunoterapia, por exemplo, que podem gerar benefícios como alta eficácia e baixos efeitos colaterais, são altamente custosos. Além disso, por consequência da demanda insuficientemente atendida, os médicos são forçados a atender o máximo de pacientes no menor tempo possível, algo que culmina em um atendimento fundamentalmente impreciso.

Outra importante fraqueza, segundo Maria Del Pilar Estevez, diretora do Corpo Clínico do ICESP, é a dificuldade de acesso ao uso de novas tecnologias, uma das principais razões para que o câncer de mama ainda seja a causa de alta mortalidade no Brasil: "A maior parte dos médicos atua no SUS e no setor privado e, com isso, vivencia situações muito díspares. Falta equidade".

Similarmente, o oncologista Stephen Stefani, presidente da ISPOR (*International Society for Pharma-coeconomics and Outcomes research*) no Brasil, relata que uma das origens do problema de acessibilidade são as distorções no sistema: "Apenas 25% das pessoas no Brasil têm acesso a planos de saúde, mas 55% dos recursos no País são gastos com essa população". Novos tratamentos, que podem custar até 10 mil dólares quando atingem o mercado, pesam ainda mais o sistema, e aumentam ainda mais a distorção. "Qualquer incorporação de um novo medicamento, se não for feita com cuidado, pode aumentar o número de excluídos. Os recursos são limitados, e não podemos conceder qualquer tipo de desperdício."

Fonte: ESTADO DE S.PAULO, 2019

### **Ameaças:**

O Ministério Público Federal (MPF) entrou com uma ação contra o Governo de São Paulo para que providências sejam tomadas a respeito da lei federal que determina que pacientes com câncer devem receber tratamento em até 60 dias após o diagnóstico. O MPF destacou que mais de 18 mil pessoas aguardam mais de dois meses entre o diagnóstico e o começo da terapia.

De acordo com a Secretaria Estadual da Saúde (SES), 1.536 pessoas continuam na espera por cirurgias para tratamento de câncer no estado e, em alguns casos, a espera chega a oito meses<sup>1</sup>.

Posto que atendimento no ICESP pré-requisita pertencimento a determinado espaço geográfico, alguns indivíduos alteram o comprovante de residência para tornarem-se elegíveis a receber tal atendimento.

A boa alimentação é um fator importante para a prevenção do câncer, manter uma dieta equilibrada pode ajudar na prevenção de diversas doenças. Alimentos ricos em fibras, vitaminas e antioxidantes oferecem inúmeros benefícios ao organismo. Portanto, a má alimentação da população pode agravar os casos de câncer.<sup>2</sup>

<sup>1</sup> (G1 SP, 2023)

<sup>2</sup> (ICESP, 2022)

### **Oportunidades:**

Segundo a OMS (Organização Mundial da Saúde, um dos principais alicerces para futuro controle do câncer é conscientizar a população para que essa sempre vise prevenir o câncer, e se cabível, detectá-lo precocemente<sup>1</sup>. Por isso, mostra-se pertinente que o ICESP expanda seus esforços na área de educação para o público geral.

<sup>1</sup> (INCA, 2021)

## **4.1.4. Planejamento Geral da Solução**

### **a) Qual é o problema a ser resolvido?**

A maior dificuldade que médicos de câncer de mama enfrentam consiste em decidir qual é o tratamento ideal para cada paciente, pois ele varia significativamente entre cada um. Desprovidos de uma métrica acurada para medir tal variância, os médicos são forçados a

recorrer à análise manual dos dados do paciente para decidir qual é o tratamento mais adequado. Posto que inerente à realização manual de tarefas é a drástica redução de eficiência e de efetividade, pode-se concluir que é necessária uma solução.

**b) Qual a solução proposta (Visão de negócios).**

A solução proposta é o desenvolvimento de um modelo preditivo que, indica qual tratamento possui maior probabilidade de sucesso - uma probabilidade que é adquirida por meio da meticulosa análise de múltiplos dados subproduto de detalhados relatórios sobre milhares de pacientes.

**c) Qual o tipo de tarefa (regressão ou classificação).**

Classificação, pois o output não é contínuo.

**d) Como a solução proposta deverá ser utilizada.**

A solução proposta deverá ser utilizada por um médico especialista no tratamento do paciente, cômico de que o modelo preditivo trata-se de, no máximo, uma *recomendação*. A decisão final sobre o melhor tratamento deve invariavelmente basear-se na análise dos dados coletados pelo médico sobre determinado paciente.

**e) Quais os benefícios trazidos pela solução proposta.**

O dilema que muitos médicos enfrentam ao tentar decidir qual tratamento é ideal para cada paciente possui sua resolução extremamente facilitada por meio da utilização do modelo. Dessa forma, o modelo aumenta as chances de que pacientes recebam o tratamento mais eficaz, e, ao permitir que médicos tomem decisões mais informadas e mais embasadas mais rapidamente, aumenta a eficiência do sistema como um todo.

**f) Qual será o critério de sucesso e qual métrica será utilizada para avaliá-lo.**

Extensão do tempo de sobrevivência é o principal fator utilizado para determinar sucesso, e índice de reincidência é utilizado como fator secundário.

## 4.1.5. Value Proposition Canvas

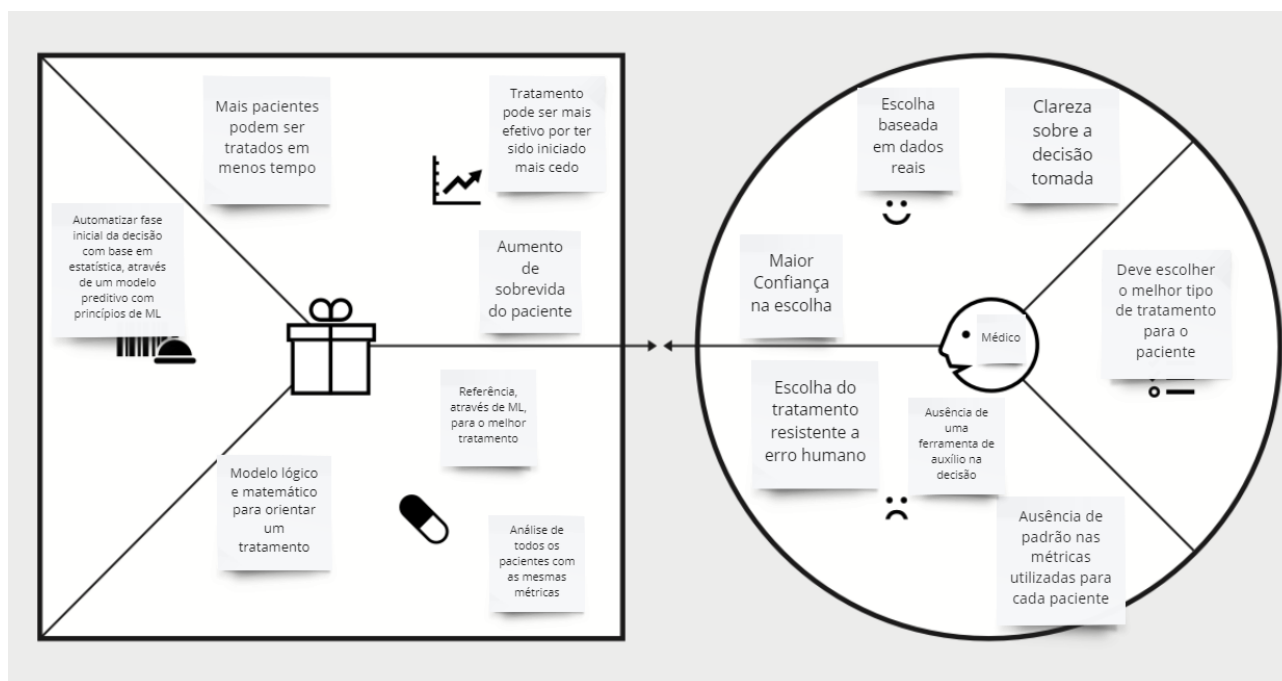


Figura 3 - Value Proposition Canvas

Fonte: Elaboração própria

## 4.1.6. Matriz de Risco

A importância da matriz de risco para o projeto reside em identificar, avaliar e priorizar os riscos potenciais, o que permite a aquisição de uma visão clara e objetiva sobre os desafios e ameaças a serem enfrentados, assim como as ações a serem tomadas para minimizar seus respectivos impactos. Além disso, a matriz de risco fornece uma base para a contínua monitorização de riscos, e, por consequência, para a constante adaptabilidade.

Figura 4 - Matriz de Risco.

Probabilidade	Ameaças					Oportunidades					Probabilidade
90%											90%
70%				F			I				70%


50%				G						50%
30%		A	D		H					30%
10%			B	C	E					10%
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
Impacto										


	NOME	CATEGORIA	PROBABILIDADE	IMPACTO
A	Descompromisso com o horário de desenvolvimento do projeto	Desenvolvimento	30%	BAIXO
B	Desentendimento entre os membros da equipe	Comunicação	10%	MODERADO
C	Médicos não fazerem a utilização do modelo preditivo	Desenvolvimento	10%	ALTO
D	Falta de dados complementares	Comunicação	30%	MODERADO
E	Falta de dados essenciais	Comunicação	10%	MUITO ALTO
F	Efeitos colaterais do tratamento	Desenvolvimento	70%	ALTO
G	Abandono do tratamento pelo paciente	Desenvolvimento	50%	ALTO
H	Os tratamentos não fazerem efeitos	Desenvolvimento	30%	MUITO ALTO
I	Ajudar na escolha do melhor tratamento para a paciente diagnosticada com câncer de mama através do modelo preditivo	Desenvolvimento	70%	ALTO

Fonte: Elaboração própria.



### 4.1.7. Personas

<b>NOME:</b>	Marcos Fernandes Neto	
<b>INFORMAÇÕES PESSOAIS:</b>		
<ul style="list-style-type: none"><li>• Marcos possui 52 anos.</li><li>• Formado em medicina.</li><li>• Trabalha como médico há 22 anos.</li><li>• É casado e possui 2 filhos.</li><li>• Nasceu e mora em São Paulo.</li></ul>		
<b>DORES:</b>	<ul style="list-style-type: none"><li>• Não sabe para qual tipo de tratamento encaminhar seus pacientes.</li><li>• Às vezes seus pacientes não voltam para dar continuidade ao tratamento.</li><li>• Muitos pacientes só procuram atendimento médico quando o câncer já está em um estágio avançado.</li></ul>	
<b>OBJETIVOS/NECESSIDADES:</b>		
<ul style="list-style-type: none"><li>• Diminuir a quantidade de casos que necessitam tratamentos mais severos e invasivos.</li><li>• Um modelo preditivo para que ele saiba para qual tratamento encaminhar o paciente.</li><li>• Deseja saber qual é o tratamento que trará mais resultados aos pacientes.</li></ul>		

<b>NOME:</b>	Renata Gonçalves Dias	
<b>INFORMAÇÕES PESSOAIS:</b>		
<ul style="list-style-type: none"> <li>• Renata possui 41 anos.</li> <li>• Está fazendo mestrado em medicina.</li> <li>• Trabalha como pesquisadora.</li> <li>• É casada.</li> <li>• Nasceu na Bahia e mora em São Paulo.</li> </ul>		
<b>DORES:</b>	<ul style="list-style-type: none"> <li>• Não sabe porque alguns pacientes respondem melhor ao tratamento neo enquanto outros respondem melhor ao tratamento adjuvante.</li> <li>• Os dados frequentemente possuem informações vazias ou são insuficientes.</li> <li>• Desconhece o que influencia na taxa de sucesso dos tratamentos.</li> </ul>	
<b>OBJETIVOS/NECESSIDADES:</b>		
<ul style="list-style-type: none"> <li>• Saber quais fatores fazem o paciente responder melhor a determinado tratamento.</li> <li>• Explorar meios que viabilizem uma diminuição dos efeitos colaterais causados pelos tratamentos.</li> <li>• Deseja descobrir se a alteração na sequência dos processos causa uma diminuição no tempo total de tratamento do paciente.</li> </ul>		

Ambas personas, em suas respectivas ocupações de médico e pesquisadora, tanto utilizarão o modelo quanto serão por ele afetadas; os resultados do modelo influencia diretamente os trabalhos que realizam, pois oferece novas lentes pelas quais o tratamento do câncer de mama pode ser visto.

#### 4.1.8. Jornadas do Usuário

O Mapa de Jornada do Usuário é uma ferramenta que ajuda a entender e acompanhar as fases pelas quais um usuário passa ao interagir com determinado modelo. Nesse caso, os usuários são médicos interagindo com plataforma que realiza análises e predições sobre qual tratamento é ideal para cada paciente.

A primeira fase é o Conhecimento da Plataforma. Nesta etapa, o médico entra em contato com o modelo pela primeira vez e precisa compreender o objetivo e o funcionamento da plataforma para que possa utilizá-la da melhor forma possível. É importante que ele saiba manusear a plataforma e entenda suas funcionalidades.

Na segunda fase, a Entrada de Dados, o médico precisa inserir informações sobre o paciente, como dados clínicos, exames, histórico médico, etc., para que o modelo possa fazer a análise e gerar uma predição. É importante que o médico preste atenção aos dados inseridos para garantir a precisão da análise.

Na terceira fase, a Saída de Informações, o médico tem acesso ao resultado gerado pelo modelo, ou seja, à predição. É importante que ele entenda o porquê da predição e que possa interpretar corretamente as informações geradas.

Por fim, a quarta e última fase é o Final do Tratamento, momento em que o médico informa a conclusão do procedimento, fornecendo dados para o modelo aprender e aumentar sua acurácia. É importante que ele preste atenção aos resultados finais para que possa contribuir para o desenvolvimento da plataforma.

Em resumo, o Mapa de Jornada do Usuário é uma ferramenta valiosa para entender e acompanhar o processo de interação do médico, nosso principal usuário, com a plataforma, garantindo que ele possa utilizá-la de maneira eficiente e eficaz.

Figura 5 - Mapa de Jornada de Usuário.

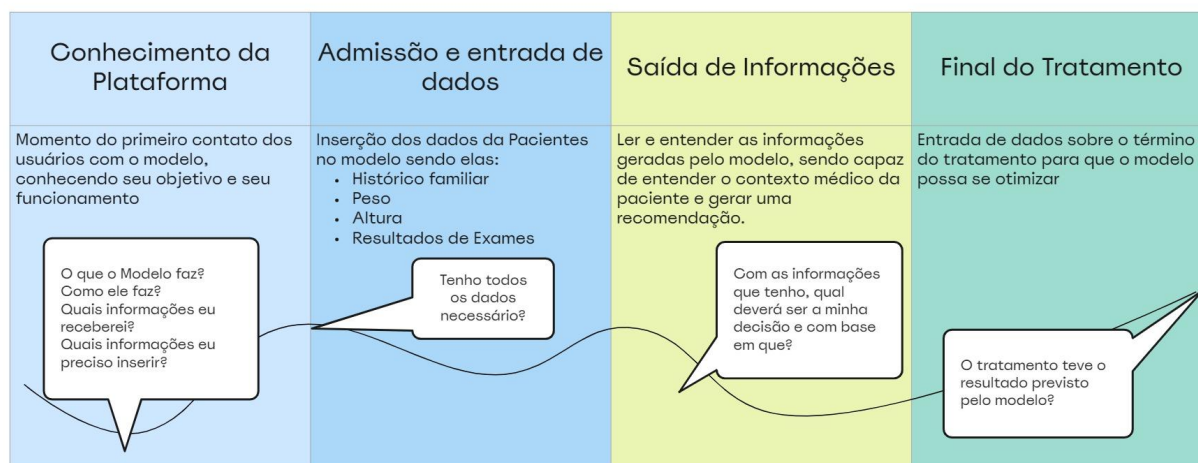


### Marcos Fernandes Neto - Médico

**Cenário:** Com o diagnóstico do paciente, ele precisa saber qual o melhor tratamento para a mesma

### Expectativas

Ele espera ser capaz de entender informações que sejam conclusivas para que ele possa indicar o tratamento mais apropriado



### Oportunidades

- Ser uma plataforma de fácil e rápido entendimento para ágil aprendizado pelos usuários
- Ter um relatório detalhado para fundamentar a lógica do modelo e facilitar o entendimento do médico sobre o por quê do resultado

### Responsabilidades

Garantir que as informações sejam claras, facilmente inseridas e entendidas pelo corpo médico

miro

Fonte: Elaboração própria.

#### 4.1.9 Política de privacidade para o projeto de acordo com a LGPD

A Pink Solution é um grupo focado na análise de dados médicos - providos do Instituto do Câncer do Estado de São Paulo (Icesp) - para a composição de um modelo preditivo para auxiliar os médicos por meio de um prognóstico de qual dos tratamentos, entre adjuvante e neoadjuvante, deveria ser recomendado ao paciente.

Nós, da Pink Solution, somos comprometidos em proteger a privacidade e segurança dos dados médicos dos nossos pacientes. Esta política de privacidade descreve como coletamos, usamos, armazenamos e compartilhamos informações pessoais sensíveis, incluindo dados médicos, na operação deste projeto de recomendação de tratamentos médicos.

##### **Coleta de Dados Médicos:**

Recebemos dados médicos dos pacientes através da base de dados do Instituto do Câncer do Estado de São Paulo (Icesp). Esses dados são coletados com o consentimento dos pacientes e são usados exclusivamente para fins médicos. As informações coletadas consistem em: idade, sexo, raça declarada, peso, altura, IMC, escolaridade, informações de estado: vivo ou óbito, informações pessoais (gravidez, menstruação, utilização de métodos contraceptivos e uso de drogas), histórico familiar de câncer e estado clínico do paciente.

##### **Uso de Dados Médicos:**

Os dados médicos coletados são usados exclusivamente para fornecer recomendações de tratamento precisas e personalizadas aos médicos, ajudando-os a tomar decisões de tratamento mais informadas para cada paciente.

##### **Armazenamento de Dados Médicos:**

Os dados médicos são armazenados em servidores seguros, protegidos por medidas de segurança físicas e digitais de alta qualidade. O acesso a esses dados é limitado a pessoal autorizado com uma necessidade legítima de conhecer essas informações, como médicos e pesquisadores. A Lei 13.787/18 disciplina a digitalização e a utilização de sistemas informatizados para a guarda, o armazenamento e o manuseio de prontuários de pacientes e o tempo de guarda dos prontuários médicos corresponde a 20 anos (MORSCH, 2022).

##### **Compartilhamento de Dados Médicos:**

Os dados médicos não serão compartilhados com terceiros, exceto quando exigido por lei ou quando houver uma necessidade médica aparente. Em tais casos, o compartilhamento será feito somente após o devido processo legal e com o consentimento dos pacientes.

**Segurança de Dados Médicos:**

Tomamos medidas rigorosas para garantir a segurança e a privacidade dos dados médicos. Isso inclui a implementação de medidas de segurança físicas e digitais, como criptografia de dados, autenticação de usuário e backup frequente.

**Direitos dos Pacientes:**

Os pacientes têm o direito de acessar, corrigir e excluir seus dados médicos a qualquer momento. Para exercer esses direitos, os pacientes devem entrar em contato através dos meios fornecidos na página de contato do Icesp.

## 4.2. Compreensão dos Dados

**1. Exploração de dados:****a) Cite quais são as colunas numéricas e categóricas.**

Primeiramente, o que são colunas numéricas ou categóricas? Coluna numérica contém valores numéricos, ou seja, valores que representam números, como, por exemplo, idade, peso, altura, temperatura, entre outros. Esses valores podem ser contínuos, quando há uma gama de valores possíveis, ou discretos, quando há valores separados e distintos.

Já uma coluna categórica contém valores que representam categorias, como, por exemplo, sexo, cor dos olhos, estado civil, entre outros. Esses valores são representados por strings ou códigos que indicam a categoria a que pertencem.

Uma forma simples de identificar se uma coluna é numérica ou categórica é observar os valores presentes na coluna. Se a maioria dos valores for números, a coluna é provavelmente numérica. Se a maioria dos valores for palavras ou frases, a coluna é provavelmente categórica.

Outra forma de identificar é utilizando funções de programação que permitam analisar os dados, como a função `describe()` no Python, que retorna um resumo estatístico de colunas

numéricas, ou a função `unique()` que retorna os valores únicos presentes em colunas categóricas.

Para reconhecer se são colunas numéricas ou colunas categóricas nós utilizamos o código abaixo para otimizarmos o tempo e conseguirmos verificar sem precisar olhar precisamente os dados. Explicando o código, caso a coluna seja igual a 'float64' ou 'int64' considera o tipo como Numérico, e se coluna igual a objeto o tipo será Categórico.

### RECONHECIMENTO DE COLUNAS NUMÉRICAS X COLUNAS CATEGÓRICAS

[+ Code](#)
[+ Markdown](#)

```
cont = 0
listacolcat = []
for coluna in tes2.dtypes:
    if coluna == 'float64' or coluna == 'int64':
        tipo = "Coluna Numérica"

    elif coluna == object:
        tipo = "Coluna Categórica"
        listacolcat.append(tes2.columns[cont])

    print(f'{tes2.columns[cont]} é {tipo}\n=====')
    cont+=1
```

Por fim, na próxima imagem teremos algumas respostas de como sairia o código acima.

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

Record ID é Coluna Numérica

=====

Idade do paciente ao primeiro diagnóstico é Coluna Numérica

=====

Última informação do paciente é Coluna Categórica

=====

### Lista de colunas numéricas

- Record ID
- Idade do paciente ao primeiro diagnóstico
- Data da última informação sobre o paciente
- Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt\_pci]
- Quantas vezes ficou grávida?
- Idade na primeira gestação
- Por quanto tempo amamentou?
- Data da cirurgia
- Data de início do tratamento quimioterapia
- Data do início Hormonoterapia adjuvante
- Data do diagnóstico
- Data de início da Radioterapia
- Grau histológico
- Subtipo tumoral
- Receptor de progesterona (quantificação %)
- Receptor de Estrogênio (quantificação %)
- Índice H (Receptor de progesterona)
- Data do tratamento
- IMC
- Data de Recidiva
- Ki67 (%)
- Data:
- Ano do diagnóstico
- Peso
- Altura (em centímetros)
- Data da primeira consulta institucional [dt\_pci]
- Código da Morfologia de acordo com o CID-O



## Lista de Colunas Categóricas

- Já ficou grávida?
- Ultima\_informacao\_paciente
- Amamentou na primeira gestação?
- Atividade Física
- Regime de Tratamento
- Tipo de terapia anti-HER2 neoadjuvante
- Radioterapia
- Esquema de hormonioterapia
- Diagnostico primario (tipo histológico)
- Receptor de estrogênio
- Receptor de progesterona
- Ki67 (>14%)
- HER2 por IHC
- HER2 por FISH
- Código da Topografia (CID-O)
- Estadio Clínico
- Grupo de Estadio Clínico
- Classificação TNM Clínico - T
- Classificação TNM Clínico - N
- Classificação TNM Clínico - M
- Combinação dos Tratamentos Realizados no Hospital
- Lateralidade do tumor
- Local de Recidiva a\xa0 distancia/ metastase #1 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #2 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #3 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #4 - CID-O - Topografia
- Com recidiva à distância
- Com recidiva regional
- Com recidiva local

## b) Estatística descritiva das colunas.

Usamos a função `describe()` o qual é um método do objeto Data Frame do Pandas, que retorna um conjunto de estatísticas descritivas para as colunas numéricas do Data Frame. Essas estatísticas incluem a contagem de valores não nulos, a média (Mean), o desvio padrão (Std), o valor mínimo (Min) e máximo (Max), o primeiro quartil (25%), a mediana (50%) e o terceiro quartil (75%).

Para as colunas que contêm dados não numéricos, a função `describe()` não é aplicável, pois essas estatísticas não têm significado para esses tipos de dados. Nesses casos, é possível usar outros métodos do Pandas, como `value_counts()`, `unique()`, `nunique()` ou `groupby()`, dependendo do que se deseja analisar.

Para fazer a estatística descritiva, nós selecionamos apenas algumas colunas numéricas e categóricas para mostrarmos de exemplo no documento.

### Numéricas

- Idade do paciente ao primeiro diagnóstico
- Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos

**ESTATÍSTICA DESCRITIVA DAS COLUNAS NUMÉRICAS**

```
tes2.describe()
```

	Record ID	Idade do paciente ao primeiro diagnóstico \
count	3726	3726
mean	49219	54
std	20989	13
min	302	22
25%	30608	45
50%	54701	54
75%	67576	63
max	82240	89

	Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci]
count	3726
mean	1501
std	842
min	25
25%	1008
50%	1301
75%	1837
max	4503

## Catégoricas

Para fazer a estatística descritiva das colunas catégoricas utilizamos o método `.value_counts()` o qual é uma função em Python que pode ser aplicada a uma série de dados. Ele retorna uma contagem de valores únicos na série e a frequência de cada valor. A saída do `.value_counts()` é uma lista com índices correspondentes aos valores únicos encontrados na série de entrada e valores correspondentes à contagem de cada valor único na série de entrada. Essa contagem é classificada em ordem decrescente de frequência. Esse método serve para entender a distribuição de valores em uma série de dados e pode ser usado para análises exploratórias de dados.

Colunas usadas para a análise descritiva:

- Última informação do paciente
- Já ficou grávida?
- Regime de Tratamento

### ESTATÍSTICA DESCRITIVA DAS COLUNAS CATEGÓRICAS

```
for coluna in listacolcat:
    print(f'Coluna:{coluna}\n\n{tes2[coluna].value_counts()}\n\n')
```

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

Coluna:Última informação do paciente

Vivo, SOE	2536
Obito por câncer	910
Vivo, com câncer	218
Óbito por outras causas, SOE	62

Name: Última informação do paciente, dtype: int64

Coluna:Já ficou grávida?

Não Informado Gravida	2787
Sim	928
Não	11

Name: Já ficou grávida?, dtype: int64

Coluna:Regime de Tratamento

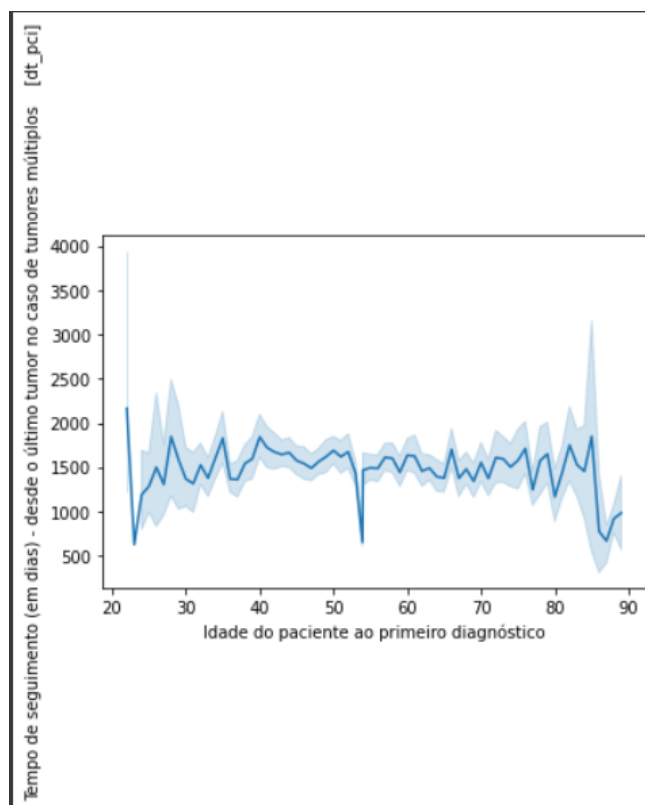
Terapia Adjuvante	1275
Não Informado Tratamento	1195
Terapia Neoadjuvante	1176
Paliativo	55

...

Name: Combinação dos Tratamentos Realizados no Hospital, dtype: int64

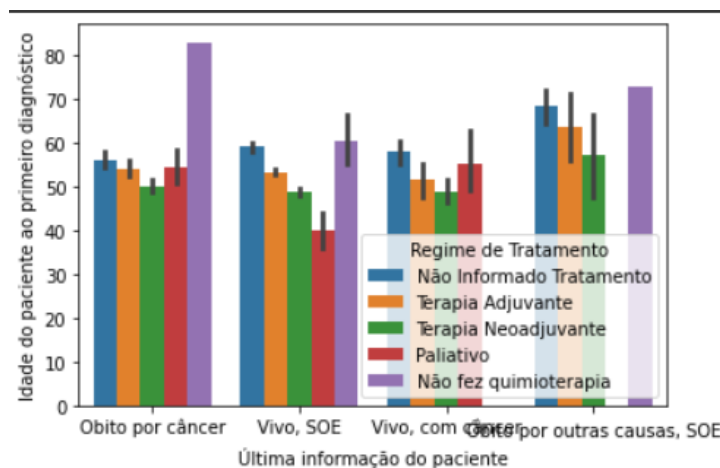
### Gráficos relacionais entre variáveis escolhidas pelo grupo

Neste primeiro gráfico relacionamos a idade do paciente e o tempo desde o último tumor.



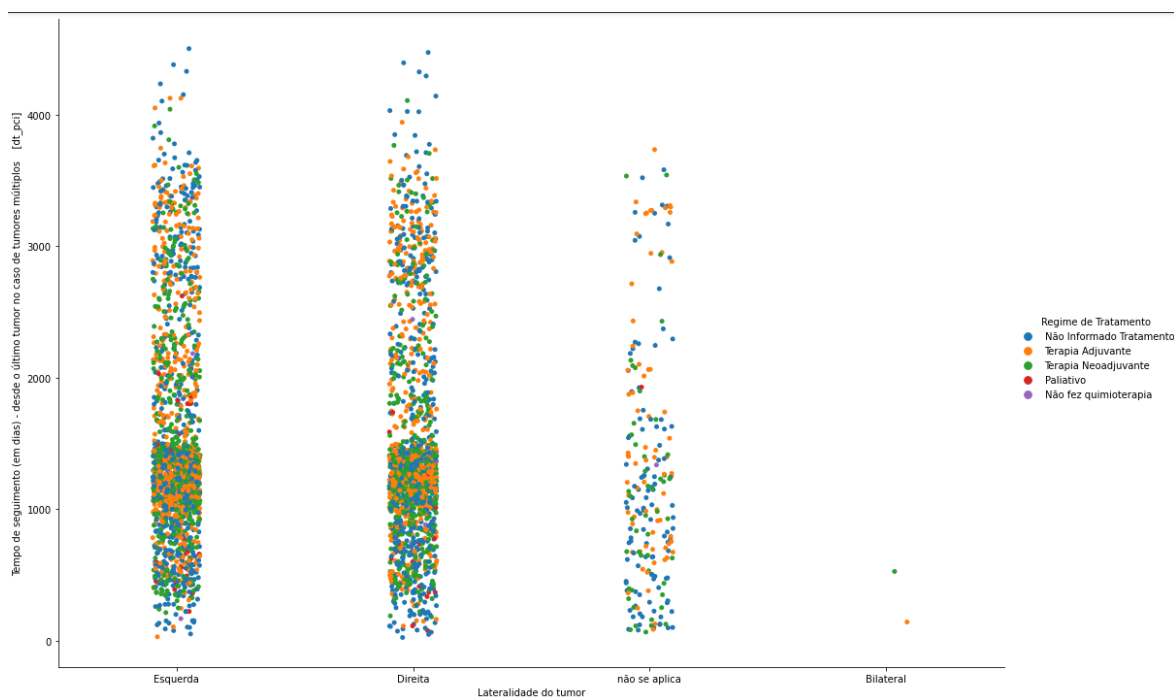
Fonte: Elaboração própria

No segundo gráfico relacionamos a situação do paciente com sua idade e explicitamos o regime de tratamento utilizado por cada um.



Fonte: Elaboração própria

Por último, este gráfico tenta enxergar a relação entre a lateralidade do tumor e o tempo desde o último tumor, novamente tentando explicitar qual foi o regime de tratamento utilizado.



Fonte: Elaboração própria

## 2. Pré-processamento dos dados:

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

Durante a análise do tratamento de dados, foram excluídos os IMCs menores que 5 e acima de 50, pois, sendo altamente improvável que um paciente alcance tal nível, a acurácia do modelo seria negativamente afetada.

Com a mesma lógica, idades acima de 90 e abaixo de 18 foram desconsiderados porque eram poucos registros e/ou não-práticos, chegando a marcar mais de 100 anos e até 0 anos. Por isso, para a construção desse modelo, foram considerados apenas adultos.

```
[ ] #SETANDO O NUMERO DE CASAS DECIMAIS
pd.set_option('display.precision',1)

menores = df_peso[(df_peso.IMC > 5)]
maiores = menores[(menores.IMC < 50)]

#Agrupando por ID e colocando a média
df_peso = maiores[(maiores.IMC != np.inf)].g

#pegar só a ultima ocorrencia
```

Na coluna "Tempo desde o ultimo tumor", retirou-se os que marcavam abaixo de 20 dias porque considerou-se que a data era muito recente.

Vale ressaltar que no pré processamento dos dados, devido ao modelo de gravação dos dados, uma mesma pessoa (Record ID) teve diversas linhas registrando suas variações de peso e altura. Para agruparmos os Record ID em uma única linha, mantivemos a última linha registrada de cada índice.

### RETIRANDO LINHAS COM MAIS DE UMA OCORRÊNCIA, MANTENDO O ÚLTIMO REGISTRO

```
[ ]

df_hist = df_hist.drop_duplicates(subset=['R
df_tumor = df_tumor.drop_duplicates(subset=[

print(df_peso['Record ID'].nunique())
print(df_hist['Record ID'].nunique())
print(df_tumor['Record ID'].nunique())
print(df_demo['Record ID'].nunique())

print('-----')

print(df_peso['Record ID'].value_counts().su
print(df_hist['Record ID'].value_counts().su
print(df_tumor['Record ID'].value_counts().s
print(df_demo['Record ID'].value_counts().su
```

### 3. Hipóteses:

Pela estratificação dos pacientes conforme a jornada de tratamento e identificação de qual estágio ele está, foi possível a identificação de três hipóteses.

- a) Levantamento das três hipóteses com justificativa.

(As hipóteses foram formuladas conforme a análise de dados com um escopo menor. Por virtude do fato de que o modelo preditivo criado foca em apenas dois tratamentos, ("Adjuvante" e "Neoadjuvante") as outras opções ("paliativo" e "não fez quimioterapia") foram descartadas; foram analisadas principalmente os padrões dos estádios clínicos mais avançados (i.e. IIIA, IIIB, IIIC e IV), pois foram os resultados que deram mais divergência entre os dois tipos de tratamento. Na coluna "última informação do paciente", foram consideradas as informações mais objetivas: "morte por câncer" e "vivo SOE". Não foram utilizadas as informações de "óbito por outras causas" pois impacta negativamente na análise e "vivo com câncer" pois existe uma série de outros fatores que podem influenciar na presença do câncer mesmo após o tratamento).

O objetivo principal é, no mínimo, prolongar a vida do paciente, e, no máximo curá-lo do câncer. Por isso, a principal métrica de sucesso utilizada foi o tempo de sobrevida do paciente.

As variáveis utilizadas são: Regime de Tratamento (Adjuvante e Neoadjuvante), última informação do paciente (as informações de "óbito por câncer" e "vivo SOE"), faixa etária\*, estágio do câncer (IIIA, IIIB, IIIC e IV), período de tratamento\*\*, contagem de Record ID e metástase\*\*\*.

### Primeira Hipótese:

Para as pessoas do grupo IIIA, o tratamento mais adequado seria o Adjuvante (independente da faixa etária).

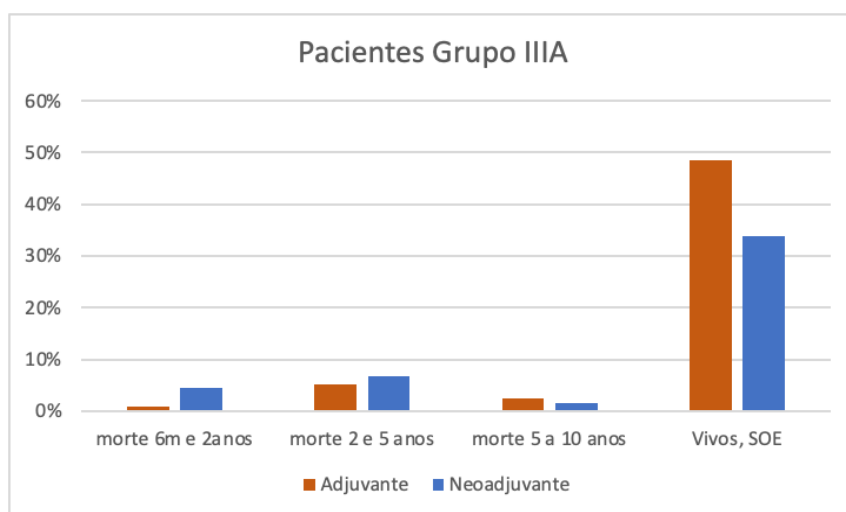


Figura 6 - Pacientes Grupo IIIA (% de mortalidade)<sup>1</sup>

Fonte: Elaboração própria.

### Segunda Hipótese:

<sup>1</sup> Para mais informações (segunda tabela da página "PIVOT ANALISE"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=120895560>



No período entre 6 meses e 2 anos, o falecimento de pacientes com até 60 anos é muito superior na terapia Neoadjuvante. Na terapia Adjuvante há uma porcentagem de falecimento de 4%, enquanto na terapia Neoadjuvante há uma porcentagem de 13%.

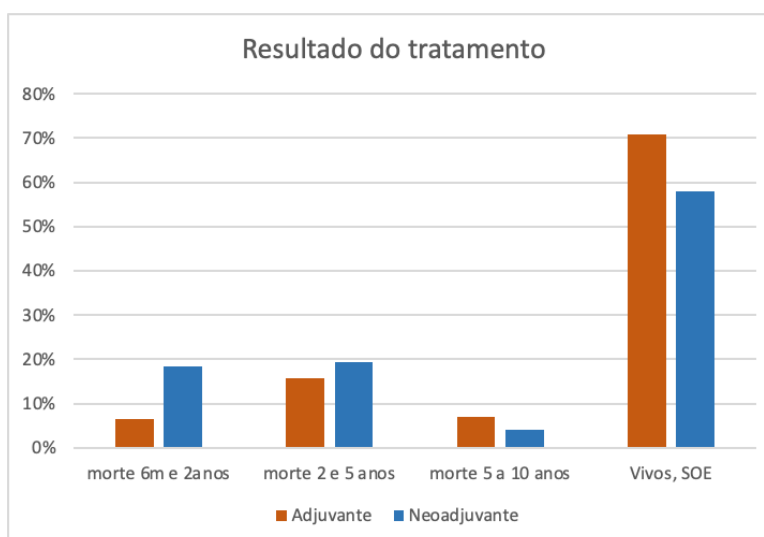


Figura 7 - Resultado do Tratamento (% de mortalidade)<sup>2</sup>

Fonte: Elaboração própria.

### Terceira Hipótese:

A diferença entre a porcentagem de ter metástase em pacientes de 40 a 60 anos é maior na terapia Neoadjuvante. Principalmente no período de 6 meses a 2 anos (Adjuvante - 0,76% ; Neoadjuvante - 2,53%).

Morte metástase	Total	40 < ID <=60
Adjuvante	4,29%	0,76%
Neoadjuvante	5,52%	2,53%

<sup>2</sup>Para mais informações (segunda tabela da página "PIVOT ANALISE"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=120895560>

Figura 8 - Mapa de Jornada de Usuário.<sup>3</sup>

*Fonte:* Elaboração própria.

\*Faixa etária: uma coluna criada a partir da informação das idades de todos os pacientes que foram subdivididos em três grupos: menos de 40 anos, entre 40 e 60 anos e mais de 60 anos.

\*\*Período de tratamento: foi calculado o período do tratamento, primeiramente em dias, por meio da subtração da "Data da Última informação do paciente" pela "Data do tratamento". Com uma coluna com o tempo do tratamento contado em dias, foi criada outra coluna que a subdividiu em cinco setores: menos de 180 dias (menos de 6 meses), entre 180 dias e 2 anos, entre 2 a 5 anos, entre 5 a 10 anos e mais de 10 anos.

\*\*\*Metástase: foi considerada a primeira coluna de presença de metástase: "Metastase ao DIAGNOSTICO - CID-O #1", para criar uma coluna de "Metastase ou não", definida pela presença de metástase ou não, independentemente de onde o câncer foi identificado no corpo.

---

<sup>3</sup> Para mais informações (tabelas da página "PIVOT Análise 2"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=909078358>

## 4.3. Preparação dos Dados e Modelagem

### 4.3.1. Modelagem para o problema

Como mencionado na Introdução, o projeto visa desenvolver inteligência artificial que, com relativo alto grau de acurácia, prevê o tratamento probabilisticamente ideal para cada paciente de câncer de mama - uma probabilidade que é adquirida por meio da meticulosa análise de múltiplos dados subproduto de detalhados relatórios sobre milhares de pacientes. Abaixo, foram detalhados os dados que possuem maior influência na escolha do tratamento.

- **Record ID:** utilizado para o tratamento dos dados - identificação de cada paciente e suas respectivas condições.
- **Idade do paciente ao primeiro diagnóstico:** é um fator importante para metrificar qual a faixa etária dos pacientes tratados. Pacientes idosos, por exemplo, podem ser inaptos ao tratamento, mas, por outro lado, pacientes mais novos podem apresentar cânceres com deterioração acelerada.
- **Última informação do paciente:** utilizada para a identificação do estado\* do paciente, é uma informação crítica para a análise e verificação de sucesso ou fracasso do tratamento.

\*tal estado pode ser "Vivo SOE"; "Vivo com câncer"; "Óbito por câncer"; "Óbito POC".

- **Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt\_pci]:** utilizado para verificar quão recente é o tumor e qual impacto teve no tratamento. Tumores mais recentes apresentam maiores chances de serem curados.
- **Já ficou grávida?:** utilizada para verificar se a gravidez influencia na escolha do tratamento, pois pacientes previamente grávidas provavelmente amamentaram, e amamentação evidencia o desenvolvimento completo da mama.
- **Regime de Tratamento:** define o tipo de tratamento escolhido (Não fez quimioterapia, Paliativo, Terapia Adjuvante e Terapia Neoadjuvante).
- **Classificação\* TNM Clínico - M:** indica a existência da presença de metástase em outros órgãos, um fator que define a complexidade do câncer. Cânceres mais complexos demandam tratamentos mais complexos.

\*As pessoas que apresentaram metástase antes do diagnóstico são indicadas para o tratamento paliativo. Portanto, nos tratamentos Adjuvante ou Neoadjuvante, significa o surgimento de metástase durante o tratamento.

- **Classificação TNM Clínico - N:** descreve se existe disseminação da doença para os linfonodos regionais; se sim, significa que o câncer começou a atacar o sistema imunológico.

- **Classificação TNM Clínico - T:** indica o tamanho do tumor primário, que influencia no estágio clínico do câncer, que pode influenciar na escolha do tratamento. (o estágio pode ser ou inicial ou avançado)
- **Lateralidade do tumor:** verifica se o local onde o tumor está localizado interfere na escolha do tratamento.
- **Com recidiva à distância, Com recidiva regional, Com recidiva local:** como mencionado na sprint 2 pela líder executiva Luciana, a presença de recidiva é um fator crítico para definir o fracasso do tratamento.
- **Estadio Clínico:** utilizado para definir e diferenciar os estádios do câncer (I, IA, IB, II, IIA, IIB, IIIA, IIIB, IIIC, IV, IVB) de cada paciente, é importante para saber qual as diferenças nas recomendações de tratamento dependendo do estágio - a recomendação de tratamentos pode variar se o estágio do câncer for avançado ou inicial.
- **Combinação dos Tratamentos Realizados no Hospital:** utilizado para verificar se outros tratamentos - radioterapia, hormonoterapia ou outras combinações - interferem na escolha do tratamento.

### 4.3.2. Métricas relacionadas ao modelo

Dentre todas as métricas apresentadas, optou-se por utilizar 4: acurácia, precisão, recall e f1-score.

A acurácia mede a quantidade de acertos nas previsões em relação a todas as previsões, ou seja, uma alta acurácia indica que há poucos casos de predições erradas, sejam elas falso positivo ou falso negativo. Tal métrica indica o quão assertivo, de modo geral, é o modelo.

Diferentemente da acurácia, a precisão indica o quão assertivo o modelo é em relação às suas predições. Isto é, indica a porcentagem de previsões corretas dentre todas as previsões "positivas". A métrica se aplica ao projeto pois retorna o grau de confiabilidade da predição de sucesso para determinado tratamento.

Por outro lado, tem-se também o Recall ou Sensibilidade que retorna a relação das previsões verdadeiras em relação a todos os casos "positivos", ou seja, retorna no modelo o quanto ele deixou de prever corretamente. Dessa forma, um recall indica que o modelo prevê um falso negativo poucas vezes - isto é, deixa de prever um caso positivo.

Por fim, ressalta-se o f1\_score, que busca retornar um equilíbrio entre a precisão e o recall, através de uma média harmônica das duas métricas. Dessa forma, o f1\_score funciona como uma métrica que indica o quão bem o modelo funciona de modo geral, mas de forma ainda mais complexa.

### 4.3.3. Apresentação do primeiro modelo candidato e discussão sobre os resultados deste modelo

Por virtude dos resultados obtidos em cada um dos métodos, e por consequência de sua lógica de funcionamento, o modelo escolhido foi o Random Forest. Inicialmente, utilizou-se os métodos K-Nearest Neighbors (KNN), Naive Bayes e Random Forest, enquanto foram feitos diversos testes de hipóteses.

A cada alteração, foram aplicados os três métodos, e comparados os resultados. Ao final, definiu-se um modelo que teve seu target criado de acordo com uma lógica baseada na coluna “Última informação do paciente”, que tinha como domínio “Vivo, SOE”, “Vivo, com câncer”, “Óbito por câncer”, “Óbito por outras causas, SOE”. A coluna target, ou seja, coluna que o modelo tenta prever, recebe apenas 0 ou 1: insucesso e sucesso, respectivamente.

Com isso em mente, foram descartadas as linhas nas quais a célula era “Óbito por outras causas, SOE”, pois, a princípio, não é possível inferir se o tratamento foi um sucesso ou não. Então, caso a célula seja “Vivo, SOE” é considerado sucesso; caso a célula seja “Vivo, com câncer” conclui-se que se o paciente estiver vivo, com câncer, e sem recidiva, consideramos sucesso; mas, se houver, recidiva, considera-se insucesso. Por fim, se a célula for “Óbito por câncer” também entende-se insucesso.

Dessa forma, o modelo é responsável por prever Sucesso ou Insucesso com base nas informações de cada paciente. Em conclusão, utilizou-se dois modelos: um responsável por prever sucesso ou insucesso de um input para o tratamento Adjuvante, e outro para prever sucesso ou insucesso para o tratamento Neoadjuvante. É importante mencionar que o modelo é idêntico, mas cada um deles foi treinado com as pacientes que tiveram o respectivo tratamento, e por isso, cada um é especializado em um tratamento.

Após os testes, foram obtidos os seguintes resultados em cada método:

#### KNN Adjuvante

Acc treino: 0.87				
Acc teste: 0.828				
	precision	recall	f1-score	support
0.0	0.53	0.20	0.30	44
1.0	0.85	0.96	0.90	206
accuracy			0.83	250
macro avg	0.69	0.58	0.60	250
weighted avg	0.79	0.83	0.80	250

### KNN Neoadjuvante

Acc treino: 0.7980241492864983					
Acc teste: 0.7017543859649122					
	precision	recall	f1-score	support	
0.0	0.57	0.51	0.54	78	
1.0	0.76	0.80	0.78	150	
accuracy			0.70	228	
macro avg	0.67	0.66	0.66	228	
weighted avg	0.70	0.70	0.70	228	

### Random Forest Adjuvante

0.885					
0.848					
	precision	recall	f1-score	support	
0.0	0.53	0.20	0.30	44	
1.0	0.85	0.96	0.90	206	
accuracy			0.83	250	
macro avg	0.69	0.58	0.60	250	
weighted avg	0.79	0.83	0.80	250	

### Random Forest Neoadjuvante

0.8309549945115258					
0.7894736842105263					
	precision	recall	f1-score	support	
0.0	0.57	0.51	0.54	78	
1.0	0.76	0.80	0.78	150	
accuracy			0.70	228	
macro avg	0.67	0.66	0.66	228	
weighted avg	0.70	0.70	0.70	228	

#### Observações:

Naive Bayes foi descartado por ter apresentado acurácia abaixo de 40%; especula-se que por consequência das colunas selecionadas.

Ambos KNN e Random Forest obtiveram acurácia maior para o tratamento adjuvante, porém o Random Forest apresenta acurácia ainda maior, sem caracterizar o overfitting. Por isso, escolheu-se o Random Forest como modelo candidato.

Random Forest, é um modelo que, como indica o nome, consiste em um conjunto de árvores de decisões: um número  $n$  de árvores com um número  $x$  de ramificações, sendo os valores de  $n$  e de  $x$  decididos pelos desenvolvedores. Tanto no Neoadjuvante quanto no Adjuvante, após escolhidas um máximo de 7 ramificações, o modelo contou as respostas das árvores e retornou aquela que mais obteve respostas segundo as árvores.

## 4.4. Comparação de Modelos

### a) Escolha da métrica e justificativa.

Utilizando como base, a partir da qualidade do modelo, os fatores de maior impacto para o problema, foi possível concluir que a métrica mais importante é a acurácia. Tal conclusão é justificada por dois fatos: a bilateralidade não-hierárquica da predição final (sucesso ou insucesso); e o alto grau de importância que possui a acurácia para avaliar-se a eficácia do modelo com relativa facilidade de interpretação (especialmente àqueles envolvidos indiretamente no processo de produção que não possuem proficiência em tecnologia), de modo que nos dê uma visão geral tanto sobre Falsos Positivos quanto Falsos Negativos..

### b) Modelos otimizados.

- Apresentar três modelos e suas métricas.

#### KNN

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV

param_grid_knn={
    'n_neighbors':[5,6,10,15,19],
    'weights':['uniform', 'distance'],
    'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'],
    'leaf_size':[10,20,30,40,50],
    'p':[1,2],
    'metric':['euclidean','manhattan','minkowski','chebyshev','mahalanobis'],
}

knn_grid_1=RandomizedSearchCV(knn,param_grid_knn,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)

knn_grid_1.fit(x_treino_knn_neoadjuvante,y_treino_knn_neoadjuvante.squeeze())
```

Métricas:

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	34
1.0	0.79	0.98	0.88	132
accuracy			0.78	166
macro avg	0.40	0.49	0.44	166
weighted avg	0.63	0.78	0.70	166

## SVC

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier

param_grid_svm={
    'C':[1,2,3,4,5,6], 'kernel':['linear', 'poly', 'rbf', 'sigmoid', 'pré-computado'], 'gamma':['scale', 'auto']}

svm_tt=svm.SVC(C=1.0)

svm_grid_neo=RandomizedSearchCV(svm_tt,param_grid_svm,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)

svm_grid_neo.fit(x_treino_svm_neoadjuvante, y_treino_svm_neoadjuvante)
```

Métricas:



	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	40
1.0	0.75	1.00	0.86	122
accuracy			0.75	162
macro avg	0.38	0.50	0.43	162
weighted avg	0.57	0.75	0.65	162

## Random Forest

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV

param_grid_rf={
    'n_estimators':[7,10,15,25,40,100,200,300,400,500], 'criterion':['gini',"entropy"],\
    'max_depth':[10,30,50,70,100], 'min_samples_split':[10,20,30,40,50,60],\
    'min_samples_leaf':[2,5,10], 'max_features':[10,30,50]
}
rf_tt=RandomForestClassifier()

rf1_grid=RandomizedSearchCV(rf_tt,param_grid_rf,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)

rf1_grid.fit(x_treino_rf_adjuvante, y_treino_rf_adjuvante)
```

Métricas:

	precision	recall	f1-score	support
0.0	0.33	0.03	0.05	34
1.0	0.80	0.98	0.88	132
accuracy			0.79	166
macro avg	0.57	0.51	0.47	166
weighted avg	0.70	0.79	0.71	166

-Os modelos apresentados foram otimizados utilizando algum algoritmo de otimização para os hiperparâmetros?

Os modelos apresentados foram otimizados utilizando Grid Search e Random Search como algoritmos de otimização para hiperparâmetros.

c) Definição do modelo escolhido e justificativa.

O modelo escolhido foi o Random Forest, por virtude do fato de que apresenta resultados superiores aos demais. O funcionamento do Random Forest consiste em construir múltiplas árvores de decisão e oferecer como output a classe selecionada pela maior parte das árvores.

a.

## 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 elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.*

## 6. Referências

G1 SP (São Paulo). **Com mais de mil pacientes com câncer à espera de cirurgia, governo de SP anuncia 45 leitos e 3 salas cirúrgicas na tentativa de reduzir fila:** segundo a secretaria estadual da saúde, meta da gestão é zerar fila nos 100 primeiros dias do ano; haverá também a ativação de 393 leitos ociosos no hospital das clínicas da faculdade de medicina da usp.. Segundo a Secretaria Estadual da Saúde, meta da gestão é zerar fila nos 100 primeiros dias do ano; haverá também a ativação de 393 leitos ociosos no Hospital das Clínicas da Faculdade de Medicina da USP.. 2023. Disponível em: <https://g1.globo.com/sp/sao-paulo/noticia/2023/01/24/com-mais-de-mil-pacientes-com-cancer-a-espera-de-cirurgia-governo-de-sp-anuncia-45-leitos-e-3-salas-cirurgicas-na-tentativa-de-reduzir-fila.ghtml>. Acesso em: 23 fev. 2023.

GIGLIO, Auro del. **ONCOLOGISTA DO HCOR APONTA 10 DICAS PARA PREVENÇÃO DO CâNCER:** a prevenção dos diversos tipos de câncer inclui, basicamente, a adoção de uma vida saudável, com alimentos que previnem o câncer e atividades físicas.. A prevenção dos diversos tipos de câncer inclui, basicamente, a adoção de uma vida saudável, com alimentos que previnem o câncer e atividades físicas.. 2021. Disponível em: [https://www.hcor.com.br/imprensa/noticias/oncologista-do-hcor-aponta-10-dicas-para-prevencao-do-cancer/?gclid=CjwKCAiAioifBhAXEiwApzCztpKeXJbn6tunOQIO8T6Cawb40AZJ6SFccPqH2riiD\\_Gx1Moi2MEvoBoCQoUQAvD\\_BwE](https://www.hcor.com.br/imprensa/noticias/oncologista-do-hcor-aponta-10-dicas-para-prevencao-do-cancer/?gclid=CjwKCAiAioifBhAXEiwApzCztpKeXJbn6tunOQIO8T6Cawb40AZJ6SFccPqH2riiD_Gx1Moi2MEvoBoCQoUQAvD_BwE). Acesso em: 23 fev. 2023.

GOVERNO, Do Portal do. **Instituto do Câncer de São Paulo recebe selo de reacreditação internacional:** icesp foi o primeiro hospital da rede pública da capital a ser acreditado pela joint commission international (jci), em 2014. Icesp foi o primeiro hospital da rede pública da capital a ser acreditado pela Joint Commission International (JCI), em 2014. 2021. Disponível em:

<https://www.saopaulo.sp.gov.br/spnoticias/orgaos-governamentais/secretaria-da-saude/instituto-do-cancer-de-sao-paulo-recebe-selo-de-reacreditacao-internacional/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **Instituto do Câncer do Estado de São Paulo**. 2022. Disponível em: <https://icesp.org.br/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **Mitos e Verdades Sobre o Câncer**. 2022. Disponível em: <https://icesp.org.br/mitos-e-verdades-sobre-o-cancer/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **RESIDENTES DA ONCOLOGIA CLÍNICA DO ICESP OBTÊM MÉDIA MAIS ALTA EM EXAME MUNDIAL**. 2023. Disponível em: <https://icesp.org.br/noticias/residentes-da-oncologia-clinica-do-icesp-obtem-media-mais-alta-em-exame-mundial/>. Acesso em: 23 fev. 2023.

INCA (Rio de Janeiro). **Detecção Precoce do Câncer**. 2021. Disponível em: <https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/deteccao-precoce-do-cancer.pdf>. Acesso em: 23 fev. 2023.

MORSCH, José Aldair. **TEMPO DE GUARDA DE PRONTUÁRIO MÉDICO: VEJA QUAL É O PRAZO E COMO SE ORGANIZAR**. 2022. Disponível em: <https://telemedicinamorsch.com.br/blog/tempo-de-guarda-de-prontuario-medico#:~:text=O%20tempo%20de%20guarda%20de%20prontu%C3%A1rio%20m%C3%A9dico%20no%20Brasil%20corresponde,Em%20seu%20Art>. Acesso em: 23 fev. 2023.

O ESTADO DE S.PAULO (São Paulo). **Acesso a novos tratamentos pelo SUS ainda é um obstáculo**: drogas mais modernas têm alto custo, e a maioria não está disponível no sistema público. Drogas mais modernas têm alto custo, e a maioria não está disponível no sistema público. 2019. Disponível em: <https://www.anahp.com.br/noticias/acesso-a-novos-tratamentos-pelo-sus-ainda-e-um-obstaculo/>. Acesso em: 23 fev. 2023.

## Anexos

*Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.*