



**G4**  
**Faculdade de**  
**Medicina da USP**

## Controle do Documento

### Histórico de revisões

Data	Autor	Versão	Resumo da atividade
01/02/2023	Luiz Fernando Covas	1.0	Criação do documento e 4.1.4 Atualização da seção 2.7
01/02/2023	Rafael Techio	1.1	Atualização 4.1.3
02/02/2023	Henrique Burle	1.2	Atualização 4.1.5
04/02/2023	Giuliano Bontempo	1.3	Atualização da seção 4.1.6
06/02/2023	Esther Hikari	1.4	Atualização da seção 4.1.2 (Análise SWOT).
06/02/2023	Felipe Moura	1.5	Preenchimento seção 4.1.7
06/02/2022	Renan Ribeiro	1.6	Introdução 4.0
06/02/2023	Renan Ribeiro	1.6	Contexto da Indústria 4.1.1
10/02/2023	Rafael Techio e Luiz Fernando Covas	1.7	Revisão geral para entrega 1
10/02/2023	Giuliano Bontempo	1.8	Revisão para a entrega 1
16/02/2023	Esther Hikari	1.9	Correção
16/02/2023	Giuliano Bontempo	1.9.1	Correção da seção 4.1.6
25/02/2023	Luiz Fernando Covas	2.0	Preenchimento das seções 4.2.1 e 4.2.2
26/02/2023	Esther Hikari Kimura Nunes	2.1	Revisão geral, adição e preenchimento da seção 4.2.4 (Política de privacidade LGPD)
27/02/2023	Esther Hikari Kimura Nunes	2.2	Preenchimento da seção 4.2.3 (Hipóteses)
27/02/2023	Giuliano Bontempo	2.3	Complemento da seção 4.2.2

# 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. 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
4.2.1. Exploração dos dados	
4.2.2. Pré-processamento	
4.2.3. Hipóteses	
4.2.4. Política de privacidade LGPD	
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 de Câncer do Estado de São Paulo (ICESP) – Octavio Frias de Oliveira, nosso parceiro neste módulo, é uma das unidades do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP), com atendimento exclusivo para pacientes da rede pública de saúde do SUS (Sistema Único de Saúde). Inaugurado em maio de 2008, o ICESSP é administrado pela Organização Social de Saúde (OSS) e a Fundação Faculdade de Medicina, por meio do Contrato de Gestão nº 01/2022, Processo HCFMUSP nº 68.919/2021.

Após mais de uma década de funcionamento, o ICESSP já atendeu mais de 121 mil pacientes do SUS, sendo que 36 mil permanecem em atendimento. A assistência com excelência é premissa básica no atendimento realizado pelo ICESSP e transcende o ato de cuidar do paciente que se encontra em suas dependências.

Nestes 14 anos desde a sua implantação, o ICESSP se consolidou como referência no atendimento oncológico do país, com elevada qualidade técnica, e desenvolvendo pesquisas e atividades de ensino em todas as áreas relacionadas à oncologia.

Entre outros benefícios garantidos por lei, o usuário do SUS tem direito a começar o tratamento do câncer – incluindo cirurgia, quimioterapia ou radioterapia –, em até 60 dias a partir da data em que foi emitido o laudo do exame que comprovou a doença.

O câncer de mama é uma doença causada pela multiplicação desordenada de células da mama. Esse processo gera células anormais que se multiplicam, formando um tumor. Visto que, a evolução do câncer de mama e sua resposta a tratamentos convencionais é muito variável. Conseguimos identificar padrões preditivos dessa variabilidade a partir de dados clínicos e do seguimento desses pacientes?

Para responder esta pergunta, foi nos proposto a criação de modelos preditivos a partir de cortes de pacientes acompanhados em projetos de pesquisa do Instituto do Câncer do Estado de São Paulo/Faculdade de Medicina da Universidade de São Paulo.

## 2. Objetivos e Justificativa

### 2.1. Objetivos

Desenvolver um modelo preditivo construído com base no dataset fornecido pelo ICESP para prever a melhor tomada de decisão em relação ao tipo de tratamento que se deve sugerir para pacientes portadores do câncer de mama, sendo eles, neoadjuvante ou adjuvante.

### 2.2. Proposta de Solução

Na área de saúde, o diagnóstico de um paciente com câncer de mama apresenta muitas variáveis que influenciam na hora de decidir qual o melhor tratamento a ser seguido. Assim, por um modelo preditivo, podemos identificar informações específicas que agrupem pacientes com base nessas variáveis genéticas, auxiliando o médico na tomada de decisão entre neoadjuvante ou adjuvante, aumentando a taxa de sucesso dos tratamentos e impactando diretamente a vida dessas pessoas.

### 2.3. Justificativa

Com a implementação do nosso modelo preditivo, poderemos ajudar os médicos na tomada de decisão de qual será o melhor tratamento para cada paciente baseado em dados passados. Assim, aprimoramos e aumentamos a taxa de efetividade na escolha. Dessa forma, conseguiremos aumentar o tempo de sobrevida dos pacientes portadores de câncer.

## 3. Metodologia

*Descreva as etapas da metodologia CRISP-DM que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.*

## 4. Desenvolvimento e Resultados

### 4.1. Compreensão do Problema

#### 4.1.1. Contexto da indústria

Um estudo feito pelo Observatório de Oncologia em 2016 revelou que o custo médio por paciente do tratamento para câncer de mama no estágio inicial fica em torno de R\$11,3 mil. No terceiro estágio, o valor médio por pessoa sobe para R\$55 mil.

Apesar do alto custo do tratamento do câncer de mama, todos os procedimentos oncológicos podem ser realizados gratuitamente pelo SUS (Sistema Único de Saúde).

Segundo dados divulgados pelo Ministério da Saúde, entre 2019 e 2021 a pasta investiu mais de R\$379 milhões para a realização de 8,7 milhões de exames de mamografia. Além disso, mais de R\$14,5 milhões foram desembolsados para 16,1 mil em reconstruções mamárias e R\$21,7 milhões foram destinados a 51,4 mil cirurgias para o tratamento do câncer. Outros R\$714 milhões foram usados para 4,2 milhões de tratamentos de radioterapia e quimioterapia, no mesmo período.

De acordo com dados da pesquisa “Número de casos e gastos com câncer de mama no Brasil atribuíveis à alimentação inadequada, excesso de peso e inatividade física”, elaborada pela Coordenação de Prevenção e Vigilância (Conprev) do INCA e apresentada em 2021, cerca de 13% dos casos de câncer de mama no Brasil ocorridos no ano passado poderiam ser evitados pela redução de fatores de risco relacionados ao estilo de vida, em especial, da inatividade física.

A pesquisa também apontou que quase 13% dos gastos federais do SUS em 2018 com o tratamento de câncer de mama (R\$102 milhões) seriam poupados pela redução de fatores de

risco comportamentais, mais uma vez com atenção especial à atividade física, que detém a maior fração (5%) dos casos de câncer de mama evitáveis pela adoção da prática.

Muitos avanços vêm ocorrendo no tratamento do câncer de mama nas últimas décadas. Há hoje mais conhecimento sobre as variadas formas de apresentação da doença e diversas terapêuticas estão disponíveis.

O tratamento do câncer de mama depende da fase em que a doença se encontra (estadiamento) e do tipo do tumor. Pode incluir cirurgia, radioterapia, quimioterapia, hormonoterapia e terapia biológica (terapia alvo).

Quando a doença é diagnosticada no início, o tratamento tem maior potencial curativo. No caso de a doença já possuir metástases (quando o câncer se espalhou para outros órgãos), o tratamento visa prolongar a sobrevida e melhorar a qualidade de vida.

O tratamento varia no estadiamento da doença, as características biológicas do tumor e as condições do paciente (idade, se já passou ou não pela menopausa, doenças preexistentes e preferências).

### **Modelo de negócio**

O ICESP é uma instituição de atendimento especializado em tratamento oncológico que segue os princípios do SUS.

O hospital atende apenas pacientes encaminhados pela rede estadual de saúde, ou seja, que foram diagnosticados com câncer em atendimentos médicos realizados nas Unidades Básicas de Saúde (UBS), Ambulatórios de Especialidades (AMES) e hospitais gerais.

O encaminhamento para o ICESP é viabilizado por meio de uma Central de Regulação de Vagas (CROSS) da Secretaria de Estado da Saúde (SES), priorizando regiões da cidade que tenham o Instituto do Câncer como referência.

O trabalho da CROSS, portanto, é garantir que os pacientes sejam encaminhados para os centros especializados em tratamento oncológico localizados próximos de sua residência, baseado em protocolos clínicos de atendimento e, em alinhamento com os fluxos de contrarreferenciamento de retorno para a região de origem.



## Concorrentes

O ICESP (Instituto do Câncer do Estado de São Paulo) tem tanto concorrentes do setor público, como também do setor privado. Hospitais como Albert Einstein, São Luiz e Sírio Libanês. A seguir alguns exemplos de competidores:

### **Hospital Israelita Albert Einstein:**

O Hospital Israelita Albert Einstein é um hospital brasileiro, privado e localizado no distrito do Morumbi, zona sul do município de São Paulo. Além de o hospital ter sido reconhecido pelo segundo ano consecutivo, o Centro de Oncologia e Hematologia Einstein alcançou a 20ª posição no ranking mundial e em relação a 2021, melhorou de posição. Um reconhecimento que ressalta o compromisso da instituição em oferecer excelência a todos os pacientes

Criado para ser o mais avançado polo de prevenção e tratamento do câncer na América Latina, o Centro de Oncologia e Hematologia Einstein Família Dayan – Daycoval combina tecnologias de última geração, recursos humanos altamente qualificados e abordagem multi e interdisciplinar para proporcionar aos pacientes cuidados completos e integrados. Do diagnóstico ao tratamento, além de uma vasta gama de serviços de suporte, como medicina integrativa, nutrologia, odontologia e cuidados paliativos, entre vários outros, o Centro congrega toda a cadeia de atendimento. Isso assegura uma abordagem holística, contemplando todas as dimensões-chave para uma assistência oncológica diferenciada, comparável à prestada nos centros de referência internacional.

O Einstein conta, ainda, com um pioneiro Centro de Medicina Personalizada. Seu laboratório de Genômica realiza sequenciamento genético e oferece um leque de cerca de 700 exames baseados em tecnologias genéticas e genômicas. São recursos que possibilitam maior precisão no diagnóstico e caracterização dos tumores, permitindo a individualização do tratamento e a identificação do medicamento mais eficaz para cada caso. A genética também é um trunfo fundamental para a detecção de risco hereditário de câncer (e de outras doenças) e para o estabelecimento de um plano de tratamento e/ou prevenção mais assertivo.

**Hospital São Luiz:**

Trata-se de um hospital privado, com sede em São Paulo e Brasília. O Centro de Oncologia São Luiz tem um corpo clínico altamente qualificado e integrado a uma equipe multiprofissional preparada, composta por psicólogos, nutricionistas, fisioterapeutas, fonoaudiólogos e enfermeiros.

Considerando que a agilidade entre a suspeita do diagnóstico e o início do tratamento são fatores decisivos para elevar o índice de cura em patologia oncológicas, foi implantado o conceito Linha Verde, visando reduzir o tempo entre o primeiro contato com o serviço, a realização dos exames, a identificação da doença e o início do tratamento.

O paciente tem um suporte completo em todas as fases e estágios da doença, incluindo quimioterapia, cirurgias oncológicas minimamente invasivas ou guiadas por robô, radiologia intervencionista, contando com a retaguarda e a segurança de um complexo hospitalar de alto padrão.

**Sírio Libanês:**

Criado em 2003, o Núcleo de Mastologia do Sírio-Libanês é formado por especialistas que aliam conhecimento e experiência, com agilidade de resolução, sofisticação tecnológica, atendimento integral multidisciplinar e, acima de tudo, atenção humanizada. O Núcleo atende mulheres e homens; crianças, adolescentes, adultos e idosos.

Os pacientes recebem assistência das equipes de Oncogenética, Oncogeriatria, Cardio-Oncologia, Cuidados Paliativos, Cuidados Integrativos e do Serviço de Voluntários do Hospital, o que confere ao atendimento um caráter integral e, em simultâneo, especializado e humanizado.

## 5 Forças de Porter

As 5 forças de Porter é o nome dado para um modelo criado por Michael Porter visando entender as forças do mercado que influenciam no desempenho de uma empresa. A seguir, uma análise do ICESP seguindo o modelo de Porter:

### **Rivalidade entre concorrentes**

Tendo em vista que o ICESP é um hospital referência no tratamento de câncer de mama, ocorre haver uma disputa por parte de pacientes, tanto de classe média baixa (que não tem condições de arcar com o tratamento) quanto de classe média alta (que mesmo tendo condições de pagar por um tratamento em um hospital particular, preferem um tratamento com maior chances de um resultado promissor, em um hospital público como o Instituto de Câncer do Estado de São Paulo). Sendo assim, o ICESP compete não só com outros hospitais públicos, mas também com o setor privado.

### **Poder de barganha dos fornecedores**

O poder de barganha dos fornecedores é relativamente baixo, pois pelo fato do ICESP se tratar de um dos hospitais de referência do país, o que o faz ter uma alta demanda de consumos para tratamentos relacionados a câncer, o hospital possui um maior poder de barganha em relação aos insumos por consumir em abundância, sendo um cliente importante para seus fornecedores e em relação à mão de obra por ser um hospital onde muitos médicos almejam trabalhar devido ao seu prestígio no mercado.

### **Poder de barganha dos compradores**

O poder de barganha dos compradores é bem baixo, o que se deve principalmente à gravidade da doença e ao fato de o ICESP ser a principal referência no tratamento de câncer. Como o câncer é uma ameaça direta à vida das pessoas afetadas por ele, elas colocam como sua prioridade máxima garantir o melhor tratamento possível, ou seja, ir na melhor instituição da sua região, que, para a maioria dos paulistanos, é o ICESP. Assim,

devido à alta infraestrutura, qualificação dos médicos e fama do ICESP, o poder de barganha dos compradores é baixo, já que ninguém barganha quando o assunto é a sua saúde.

#### **Ameaça de novos entrantes:**

A ameaça de novos entrantes, como outros hospitais públicos, se qualificarem e se tornarem referência no tratamento do câncer, ou que novos hospitais particulares possam surgir é relativamente baixa. Isso ocorre por se tratar de um tratamento muito complexo, que necessita de muito investimento. Dificilmente haverá algum hospital com tanto conhecimento sobre o assunto e com o nível de experiência que o ICESP possui.

#### **Ameaça de produtos ou serviços substitutos**

Podemos considerar baixa a ameaça de substitutos por se tratar de um assunto muito complexo, o qual demanda muito estudo, pesquisa e investimento. Sendo assim, a probabilidade de surgir um novo tratamento como uma possível cura ou algo nesse sentido é reduzida.

### 4.1.2. Análise SWOT

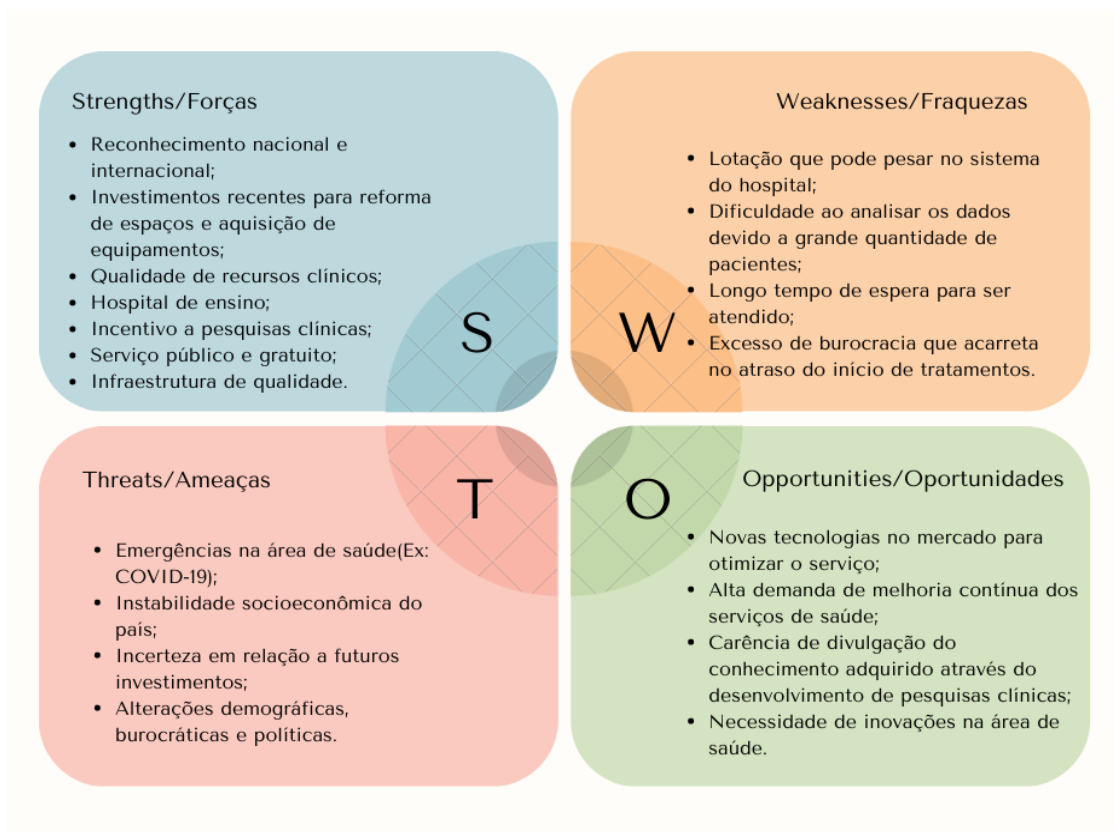


Imagem 1: Canva Análise SWOT.

Fonte: Desenvolvido pelo próprio grupo através do canva.com

### 4.1.3. Planejamento Geral da Solução

#### 3.1) Qual é o problema a ser resolvido

Dificuldade na tomada de decisão do melhor tratamento e na análise dos dados disponíveis.

#### 3.2) Qual a solução proposta (visão de negócios)

Desenvolver um modelo preditivo com o intuito de auxiliar médicos a encontrar de forma mais rápida e assertiva a melhor opção de tratamento entre a abordagem neoadjuvante (quimioterapia seguida de cirurgia) e adjuvante (cirurgia seguida de quimioterapia) nos casos de câncer de mama.

### **3.3) Como a solução proposta deverá ser utilizada**

A solução será implementada em um site web de modo que o médico poderá selecionar características clínicas da paciente e assim obter como resultado o tratamento ideal.

Futuramente, o algoritmo de modelo preditivo poderá ser implementado no sistema hospitalar de modo a automatizar o processo de consulta ou até mesmo estar inserido em uma API para poder ser acessado de vários outros sistemas.

### **3.4) Quais os benefícios trazidos pela solução proposta**

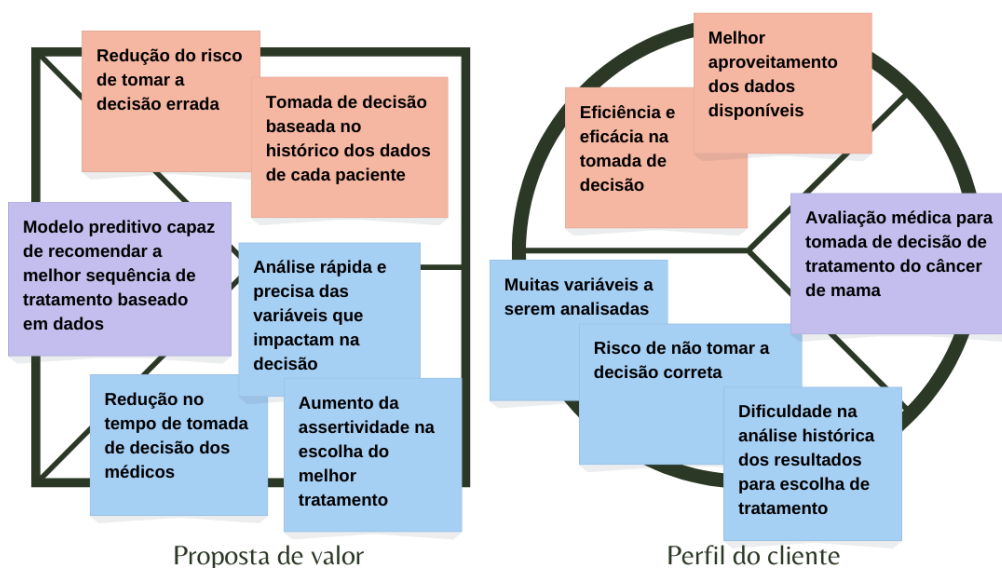
O auxílio ao médico responsável por definir qual tratamento o paciente deverá seguir com maior assertividade e um possível impacto na cura do paciente, além de um melhor aproveitamento dos dados disponíveis.

### **3.5) Qual será o critério de sucesso e qual medida será utilizada para o avaliar**

O critério de sucesso será definido a partir da taxa de assertividade dos testes realizados. Os testes do algoritmo serão avaliados comparando suas respostas às análises de casos disponibilizadas pela USP, considerando todas as variáveis genéticas existentes nestes tratamentos e entregando um resultado que seja conciso com o tipo de tumor do paciente, assim como outros fatores.

#### 4.1.4. Value Proposition Canvas

### Value Proposition Canvas



**Imagem 2:** Value Proposition Canvas.

**Fonte:** Desenvolvido pelo próprio grupo através do Canva.com.

#### 4.1.5. Matriz de Riscos

Probabilidade	Ameaças					Oportunidades					Possibilidade
90%						Ajuda no combate ao câncer de mama	Aprendizado para membros do grupo				90%
70%			Tarefas mal divididas	Problemas com dados							70%
50%					O modelo não atender as demandas						50%
30%				Não entender os dados	Instabilidade de algoritmos						30%
10%			Faltar engajamento dos participantes	Grupo se desentender							10%
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo	

**Imagem 3:** Matriz de risco.

**Fonte:** Desenvolvido pelo próprio grupo através do Excel.

## 4.1.6. Personas

Marco Aurélio - Utiliza o modelo

### Marco Aurélio



*"O melhor médico é aquele que mais esperança inspira"*

Idade: 36  
Profissão: Médico oncologista especializado em câncer de mama  
Família: **Casado, sem filhos**  
Localização: **São Paulo, SP**

**Dedicado** **Altruista** **Inteligente** **Esforçado**

#### Objetivos

- Salvar vidas
- Tornar-se referência na área de câncer de mama
- Construir uma família

#### Frustrações

- Diagnósticos passados que não foram precisos
- Pressão dada a importância do seu trabalho
- Acaba não tendo tempo suficiente para a sua esposa devido a ter muito trabalho

#### Bio

Marco Aurélio é nascido em São Paulo capital e desde pequeno sempre foi muito esforçado e considerado brilhante. Kursou medicina na USP e resolveu se especializar em oncologia, por câncer ser uma das doenças que mais mata, e ele ter o desejo de salvar vidas. Hoje ele se orgulha muito do seu trabalho e de todas as vidas que já salvou, mas reconhece que a pressão é muito alta e gostaria de ter mais ferramentas para auxiliá-lo nos diagnósticos e decisões de caminho de tratamento.

#### Motivação

Motivador	Nível
Incentivo	20%
Medo	10%
Crescimento pessoal	70%
Poder	60%
Altruismo	80%

#### Personalidade

Dimensão	Nível
Introverso	10%
Extroverso	90%
Racional	10%
Emocional	90%

Imagem 4: Persona 1.

Maria Helena - Afetada pelo modelo

### Maria Helena



*"Sofremos muito pelo pouco que nos falta e aproveitamos muito o pouco que temos"*

Idade: 51  
Profissão: Florista  
Família: **Casada, 2 filhos**  
Localização: **São Paulo, SP**

**Auto astral** **Culta** **Generosa** **Comunicativa**

#### Objetivos

- Envelhecer de maneira saudável
- Depois de aposentada morar no interior
- Se curar do câncer o mais rápido possível

#### Frustrações

- A velocidade da vida na cidade grande
- Não ter aprendido um segundo idioma quando jovem
- Medo do câncer de mama inviabilizar seus planos

#### Bio

Maria Helena sempre foi apaixonada pela vida no campo. Com o objetivo de tentar trazer um pouco disso ao seu cotidiano, ela abriu uma floricultura na cidade. Recentemente, já nos seus 50 anos de idade, descobriu que está com câncer de mama, o que a faz viver constantemente preocupada com a doença. Ela deseja se livrar do câncer o mais rápido o possível, para poder concluir seus planos de se mudar para o interior com seu marido.

#### Motivação

Motivador	Nível
Incentivo	30%
Medo	70%
Crescimento pessoal	60%
Poder	10%
Altruismo	50%

#### Personalidade

Dimensão	Nível
Introverso	10%
Extroverso	90%
Racional	10%
Emocional	90%

Imagem 5: Persona 2.



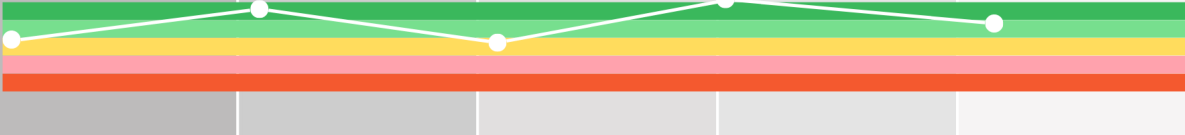
### 4.1.7. Jornadas do Usuário

Uma jornada do usuário é uma visualização do processo pelo qual uma pessoa passa para atingir um objetivo. É essencial para que toda a equipe tenha uma percepção comum sobre o processo, motivações, ações, sentimentos, pensamentos e expectativas do usuário.

**Usuário:** Marco Aurélio

**Cenário:** Marco Aurélio é um médico oncologista do ICESP que trabalha com diversos casos de câncer de mama. No hospital suas atividades são o diagnóstico, escolha do tratamento e acompanhamento dos seus pacientes. Hoje, o hospital tem uma base de dados enorme que mostra os resultados dos pacientes conforme a escolha do tratamento, o problema é que esses dados não são passíveis de uma análise preditiva, por falta de tratamento de dados e um modelo fidedigno.

**Expectativas:** Melhorar a eficiência, eficácia e facilidade na escolha do tratamento ideal, tendo como suporte adicional uma análise preditiva dos dados históricos dos tratamentos realizados no hospital. Para que se possa ter mais um elemento respaldando suas escolhas.

	Diagnóstico	Análise dos Dados	Retorno	Realizar Tratamento Neo ou Adjuvante	Acompanhamento
Atividades do Usuário	1) Triagem 2) Primeira consulta com paciente 3) Coleta dos dados	1) Análise clínica 2) Análise cirúrgica 3) Análise preditiva 4) Análise situacional	1) Informar ao paciente a gravidade dos seus exames 2) Indicar tratamentos possíveis	1) Realizar cirurgia ou quimioterapia	1) Consultas com pacientes; 2) Análise dos exames para ver o desenvolvimento do tumor; 3) Reavaliação do tratamento escolhido.
Pensamento	Será que esse câncer está muito avançado?	Acho que conseguimos curá-los!!	Será que o paciente vai voltar?	Vamos acabar com esse câncer!	Como será que está o tratamento?
Sentimento					

**Imagem 6:** Jornada do usuário.

**Fonte:** Desenvolvido pelo próprio grupo através do Canva.com

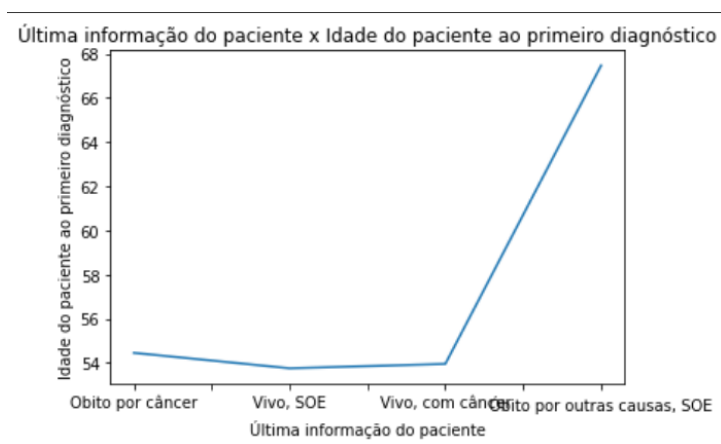
**Oportunidades:** Facilitar o preenchimento e uso do modelo, além de melhorar a visualização dos dados. Criar uma métrica por pesquisas quantitativas sobre a experiência do usuário com o modelo.

## 4.2. Compreensão dos Dados

### 4.2.1. Exploração de dados:

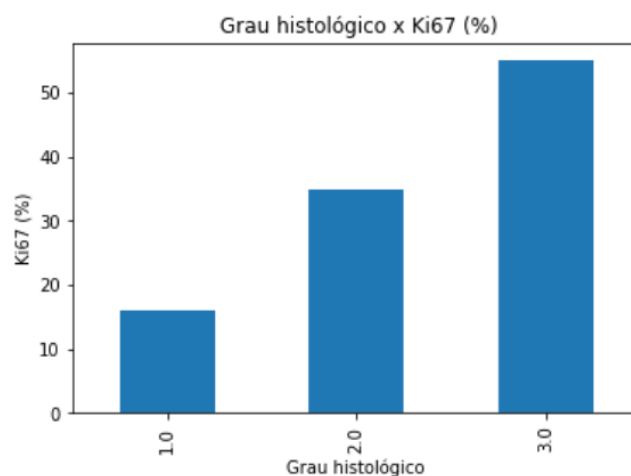
Após recebermos as quatro tabelas fornecidas pelo ICESP (Instituto do Câncer do Estado de São Paulo) com diversas informações, coletadas de prontuários médicos, a respeito de pacientes portadoras do câncer de mama, começamos a explorar os dados a fim de entender e descobrir possíveis correlações entre as variáveis. Nesse processo levantamos algumas hipóteses, que nos ajudarão na criação do nosso modelo preditivo. Para isso, utilizamos algumas ferramentas como Colab, funções do python e algumas bibliotecas, para visualizarmos e entendermos melhor como estava estruturada a nossa base de dados.

Segue abaixo alguns gráficos que plotamos em nosso notebook para visualizarmos melhor nossos dados e as variáveis disponíveis:



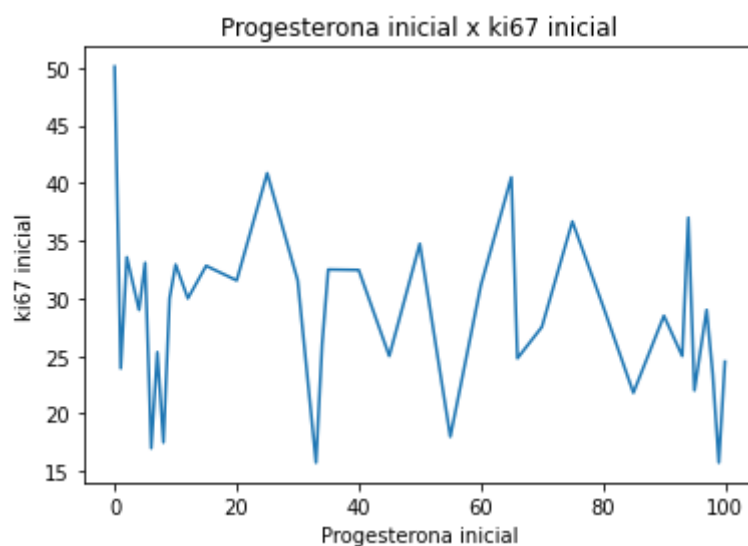
**Imagem 7:** Gráfico relacionando os tipos de tratamento e a quantidade média de dias até a primeira recidiva.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.



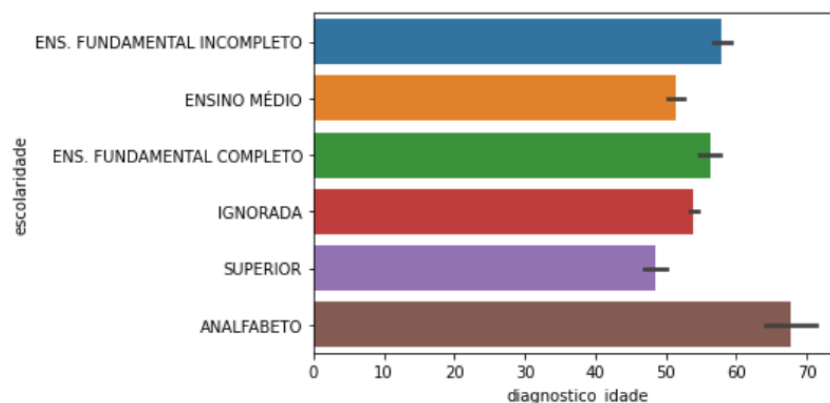
**Imagem 8:** Gráfico relacionando o grau histológico e a quantidade de Ki67, substância liberada durante a divisão celular, presente no organismo do paciente.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.



**Imagem 9:** Gráfico relacionando o valor da progesterona inicial com a quantidade de Ki67 presente no organismo do paciente.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.



**Imagem 10:** Gráfico relacionando o nível de escolaridade dos pacientes com o tempo para o diagnóstico de câncer.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

a) Cite quais são as colunas numéricas e categóricas.

Para realizar a identificação da quantidade de colunas numéricas e categóricas que nosso DataFrame **apresentava utilizamos** o método `df.info()`. Assim conseguimos obter o número total de linhas e colunas, o nome de cada coluna, o número de valores não nulos e o tipo de dados de cada coluna.

```
df_peso_altura.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55654 entries, 0 to 55653
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Record ID              55654 non-null  int64
1   Repeat Instrument       51382 non-null  object
2   Repeat Instance        51382 non-null  float64
3   Data:                  51354 non-null  object
4   Peso                   45178 non-null  float64
5   Altura (em centímetros) 49928 non-null  float64
6   IMC                    51334 non-null  float64
dtypes: float64(4), int64(1), object(2)
memory usage: 3.0+ MB
```

**Imagem 7:** Método `df.info()` sendo utilizada no DataFrame peso e altura.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

### Referente a tabela Demográficos temos:

*Colunas numéricas:*

Record ID;

Repeat Instrument ;

Repeat Instance;

Idade do paciente ao primeiro diagnóstico;

Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt\_pci];

Quantas vezes ficou grávida?;

Número de partos;

Idade na primeira gestação;

Por quanto tempo amamentou?;

Idade da primeira menstruação.

#### *Colunas categóricas:*

Escolaridade;

Sexo;

Raça declarada (Biobanco);

UF de nascimento do paciente;

UF de residência do paciente;

Data da última informação sobre o paciente;

Última informação do paciente;

Já ficou grávida?;

Abortou;

Amamentou na primeira gestação?;

Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Não);

Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 1º grau, apenas 1 caso);

Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 1º grau, mais de 1 caso);

Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, apenas 1 caso);

Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, mais de 1 caso);

Faz uso de métodos contraceptivo?;

Qual método? (choice=Pílula anticoncepcional);

Qual método? (choice=DIU);

Qual método? (choice=camisinha);

Qual método? (choice=outros);

Qual método? (choice=não informou);

Já fez uso de drogas?;

Atividade Física;

Consumo de tabaco;

Consumo de álcool;

Possui histórico familiar de câncer?;

Grau de parentesco de familiar com cancer? (choice=primeiro (pais, irmãos, filhos));

Grau de parentesco de familiar com cancer? (choice=segundo (avós, tios e netos));

Grau de parentesco de familiar com cancer? (choice=terceiro (bisavós, tio avós, primos, sobrinhos));

Regime de Tratamento;

Hormonioterapia;

Data da cirurgia;

Tipo de terapia anti-HER2 neoadjuvante;

Radioterapia;

Data de início do tratamento quimioterapia;

Esquema de hormonioterapia;

Data do início Hormonioterapia adjuvante;

Data de início da Radioterapia.

### **Referente a tabela Histopatologia temos:**

*Colunas numéricas:*

Record ID;

Repeat Instance;

Grau histológico;

Subtipo tumoral;

Índice H (Receptor de progesterona);

Ki67 (%).

*Colunas categóricas:*

Repeat Instrument;

Diagnostico primario (tipo histológico);

Receptor de estrogênio;

Receptor de progesterona;

Ki67 (>14%);

Receptor de progesterona (quantificação %);

Receptor de Estrogênio (quantificação %);

HER2 por IHC;

HER2 por FISH.

### **Referente a tabela Peso e Altura temos:**

*Colunas numéricas:*

Record ID;

Repeat Instance;

Peso;

Altura (em centímetros);

IMC.

*Colunas categóricas:*

Repeat Instrument;

Data.

## Referente a tabela Registro de tumor temos:

### *Colunas numéricas:*

Record ID;  
Repeat Instance;  
Código da Morfologia de acordo com o CID-O;  
Ano do diagnóstico;  
Tempo desde o diagnóstico até a primeira recidiva.

### *Colunas categóricas:*

Repeat Instrument;  
Data da primeira consulta institucional [dt\_pci];  
Data do diagnóstico;  
Código da Topografia (CID-O);  
Estadio Clínico;  
Grupo de Estadio Clínico;  
Classificação TNM Clínico - T;  
Classificação TNM Clínico - N;  
Classificação TNM Clínico - M;  
Metastase ao DIAGNÓSTICO - CID-O #1;  
Metastase ao DIAGNÓSTICO - CID-O #2;  
Metastase ao DIAGNÓSTICO - CID-O #3;  
Metastase ao DIAGNÓSTICO - CID-O #4;  
Data do tratamento;  
Combinação dos Tratamentos Realizados no Hospital;  
Lateralidade do tumor;  
Data de Recidiva;  
Local de Recidiva a distancia/ metastase #1 - CID-O - Topografia;  
Local de Recidiva a distancia/ metástase #2 - CID-O - Topografia;  
Local de Recidiva a distancia/ metástase #3 - CID-O - Topografia;  
Local de Recidiva a distancia/ metástase #4 - CID-O - Topografia;  
Descrição da Morfologia de acordo com o CID-O (CID-O - 3ª edição);  
Descrição da Topografia;  
Classificação TNM Patológico - N;  
Classificação TNM Patológico - T;  
Com recidiva à distância;  
Com recidiva regional;  
Com recidiva local.

## b) Estatística descritiva das colunas.

Utilizamos o método `df.describe()` do pacote `pandas` para gerar um conjunto de estatísticas descritivas dos nossos DataFrames, onde conseguimos informações

extremamente relevantes para entendermos como está o nosso DataSet permitindo a fácil identificação de valores extremos, a distribuição geral dos dados e outras informações importantes.

Exemplo de utilização na tabela Demográficos:

df\_Demograficos.describe()

	Record ID	Repeat Instrument	Repeat Instance	Idade do paciente ao primeiro diagnóstico	Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci]	Quantas vezes ficou grávida?	Número de partos	Idade na primeira gestação	Por quanto tempo amamentou?	Idade da primeira menstruação
count	4272.000000	0.0	0.0	4092.000000	4270.000000	44.000000	2.000000	897.000000	688.000000	1025.000000
mean	48652.360487	NaN	NaN	54.247801	1475.003747	2.318182	1.500000	23.057971	19.043605	12.891707
std	20659.519622	NaN	NaN	13.574088	859.622377	1.410471	0.707107	5.665232	23.105060	2.104446
min	302.000000	NaN	NaN	22.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	31013.000000	NaN	NaN	45.000000	956.250000	1.000000	1.250000	19.000000	6.000000	12.000000
50%	53394.000000	NaN	NaN	54.000000	1282.000000	2.000000	1.500000	22.000000	12.000000	13.000000
75%	65816.750000	NaN	NaN	64.000000	1817.750000	3.000000	1.750000	26.000000	24.000000	14.000000
max	82240.000000	NaN	NaN	98.000000	4503.000000	7.000000	2.000000	53.000000	260.000000	37.000000

**Imagem 8:** Método df.describe() sendo utilizado no DataFrame Demográficos.

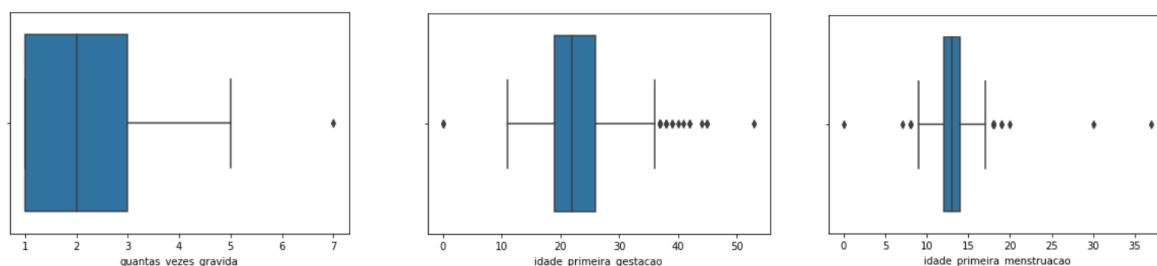
**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

#### 4.2.2. Pré-processamento dos dados:

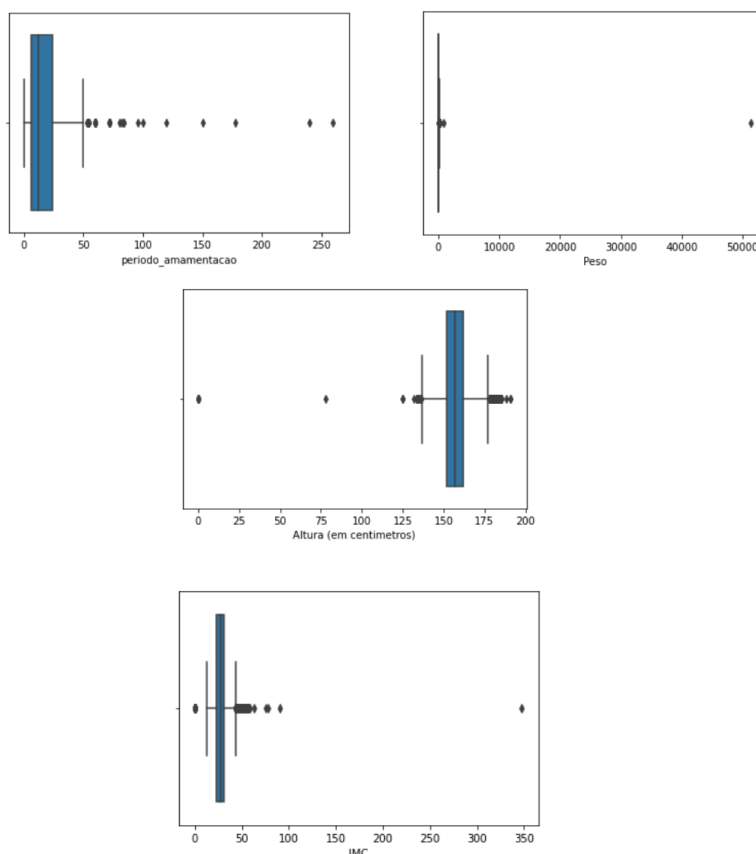
Para realizarmos o pré-processamento excluímos algumas colunas que identificamos estarem sem dados preenchidos ou com pouquíssimos dados após a exploração, a fim de não enviesar o nosso modelo com o preenchimento de muitas informações, após feito isso, partimos para a identificação dos outliers utilizando o boxplot e substituindo os valores por NaN. Na sequência começamos a tratar os missings numbers, onde para cada caso demos um peso diferente, algumas colunas utilizamos a média (tabela peso e altura), outras colunas utilizamos cálculos feitos por um intervalo da distribuição normal (tabela demográficos) e também fizemos combinações de variáveis para encontrar padrões como nas colunas da tabela Histopatologia. Por fim realizamos a codificação das colunas categóricas para uma melhor análise do modelo.

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

Exemplo da aplicação do boxplot para identificação de outliers:







**Imagem 9:** Gráficos representando outliers.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

Após identificarmos os outliers através do **boxplot** utilizamos a função abaixo para **corrigi-los substituindo-os** por um **valor nulo** na tabela Demográficos.

```
[72] numerical_cols = ['diagnostico_idade',
                      'tempo_de_seguimento',
                      'quantas_vezes_gravida',
                      'idade_primeira_gestacao',
                      'periodo_amamentacao',
                      'idade_primeira_menstruacao']

def exclui_outliers(DataFrame, col_name):
    intervalo = 3*DataFrame[col_name].std()
    media = DataFrame[col_name].mean()
    DataFrame.loc[df_Demograficos[col_name] < (media - intervalo), col_name] = np.nan
    DataFrame.loc[df_Demograficos[col_name] > (media + intervalo), col_name] = np.nan

for col in numerical_cols:
    exclui_outliers(df_Demograficos, col)
```

**Imagem 10:** Função `excluir_outliers`.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

Já para os outliers encontrados na tabela **peso e altura** utilizamos uma função que se aproveita do fato de termos várias entradas repetidas com o mesmo ID para lidar com os

outliers e NaNs. Ela divide o dataframe em subgrupos com o mesmo ID e utiliza a **mediana** das colunas deste subgrupo para substituir os valores de outliers e números faltantes.

```
contagem = 0
def tira_na(id, df):
    global contagem
    # Trata a coluna de altura
    if math.isnan(df.iloc[contagem, 5]) or df.iloc[contagem, 5] == 0:
        (df.loc[contagem, "Altura (em centímetros)"]) = df[(df["Record ID"] == id)][("Altura (em centímetros)").median()
    if math.isnan(df.loc[contagem, "Altura (em centímetros)"]):
        (df.loc[contagem, "Altura (em centímetros)"]) = df["Altura (em centímetros)"].median()
    # Trata a coluna de peso
    if math.isnan(df.loc[contagem, "Peso"]) or df.loc[contagem, "Peso"] >= 150 or df.loc[contagem, "Peso"] <= 20:
        (df.loc[contagem, "Peso"]) = df[(df["Record ID"] == id)][("Peso").median()
    if math.isnan(df.loc[contagem, "Peso"]) or df.loc[contagem, "Peso"] <= 20:
        (df.iloc[contagem, 4]) = df["Peso"].median()
    # Trata a coluna de IMC
    if math.isnan(df.loc[contagem, "IMC"]) or df.loc[contagem, "IMC"] >= 50 or df.loc[contagem, "IMC"] <= 10:
        (df.loc[contagem, "IMC"]) = (df.loc[contagem, "Peso"]) / ((df.iloc[contagem, 5] / 100) ** 2).round(1)
    contagem += 1
df_peso_altura["Record ID"].apply(tira_na, args=(df_peso_altura,))
```

**Imagem 11:** Função tira\_na.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

#### b) Normalização das colunas numéricas

Para normalizar as colunas utilizamos o módulo preprocessing da biblioteca sklearn. Mais especificamente, a sua classe MinMaxScaler. Alimentamos ele com as colunas numéricas e ele as normalizou.

Como ainda não separamos o dataframe em treino e teste, fizemos uma normalização geral. Futuramente isso será corrigido e o MinMaxScaler será treinado somente com o conjunto de treino.

Segue o código:

```
# Normalizando os dados numéricos

"""Como ainda não separamos em treino e teste, fizemos uma normalização geral.
Futuramente isso será corrigido e o MinMaxScaler será treinado somente com o
conjunto de treino."""

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

numerical_cols = df.select_dtypes(include='number').columns

# Removendo o record_id da normalização
numerical_cols = numerical_cols.drop('record_id')

df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
df.head(3)
```

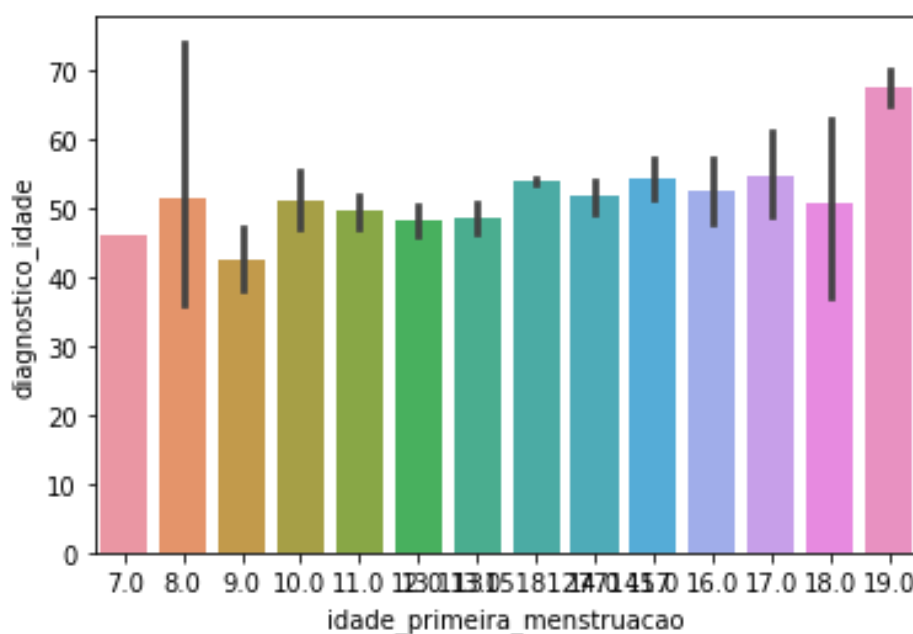
**Imagem 12:** Normalização das colunas numéricas.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

### 4.2.3. Hipóteses

a) Levantamento das três hipóteses com justificativa.

**Hipótese 1:** A idade na primeira menstruação é um fator de risco para câncer de mama?

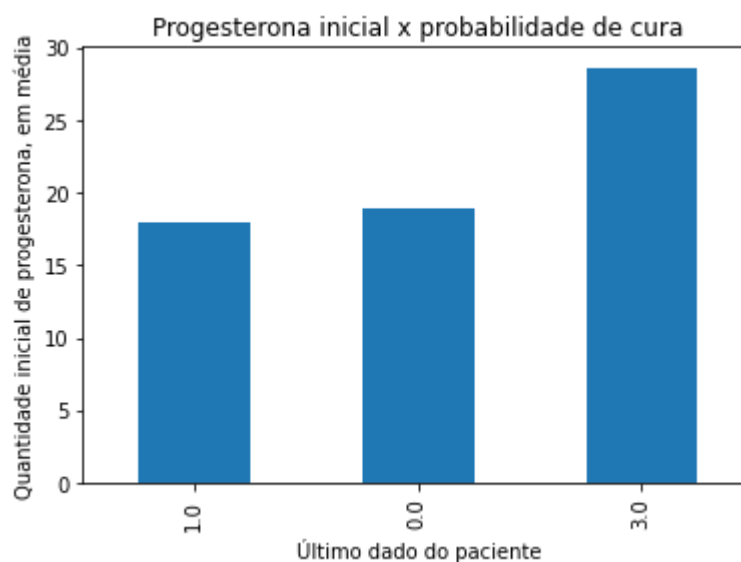


**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

**Imagem 12:** Gráfico relacionando a idade na menarca com a idade no diagnóstico.

**Justificativa:** Com base no banco de dados disponível, desenvolvemos um gráfico que mostra a idade da primeira menstruação e a idade do diagnóstico, isso possibilitou observarmos uma relação entre os acontecimentos e questionar o impacto desta na descoberta do câncer de mama nos pacientes.

**Hipótese 2:** A progesterona inicial afeta a possibilidade de cura de um paciente?

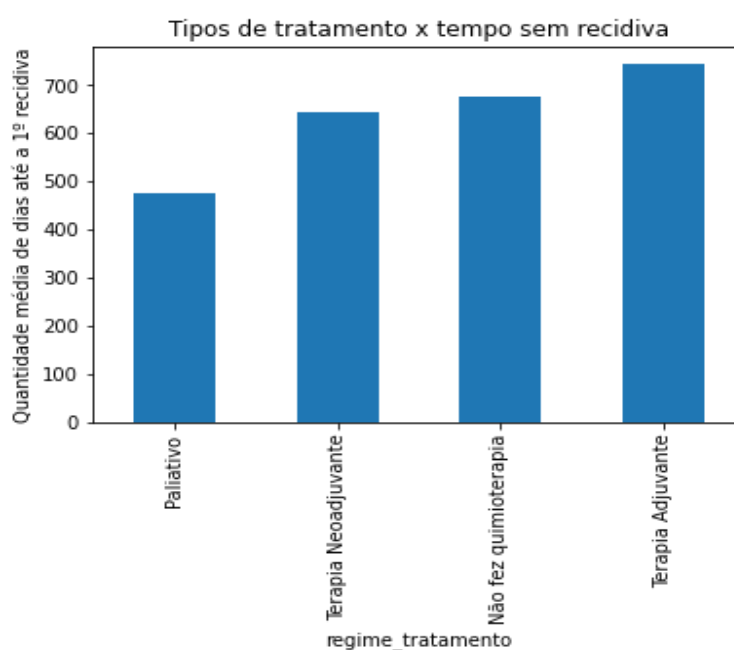


**Imagem 13:** Gráfico relacionando a quantidade inicial de progesterona com o último dado do paciente.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

**Justificativa:** Durante a exploração dos dados, notamos um padrão no valor da progesterona no primeiro exame após o início do tratamento e o tempo entre o diagnóstico e a cura. Concluímos que essa relação pode impactar nas decisões e no procedimento que deverá ser seguido para o tratamento.

**Hipótese 3: Existem grandes diferenças nos tempos sem recidiva de diferentes tratamentos?**



**Imagem 14:** Gráfico relacionando os tipos de tratamento e a quantidade média de dias até a primeira recidiva.

**Fonte:** Desenvolvido pelo próprio grupo através do Google Colab.

**Justificativa:** O gráfico acima foi desenvolvido com base no banco de dados entregue pelo ICESP. Nele, observamos a relação entre o tipo de tratamento e o tempo sem recidiva. Concluimos que o tratamento escolhido influencia na sobrevida sem câncer.

#### 4.2.4. Política de privacidade LGPD

Este modelo preditivo é mantido e operado pelo grupo G4. Grupo, que está a desenvolver um modelo preditivo com objetivo de auxiliar os médicos no tratamento do câncer de mama.

Nós coletamos e utilizamos dados pessoais pertencentes aos pacientes que concordaram em compartilhar seus dados para o treinamento do nosso modelo. Ao fazê-lo, agimos na qualidade de **controlador** desses dados e estamos sujeitos às disposições da Lei Federal n. 13.709/2018 (Lei Geral de Proteção de Dados Pessoais - LGPD).

Nós cuidamos da proteção de seus dados pessoais e, por isso, disponibilizamos esta política de privacidade, que contém importantes informações.

##### 1. Bases legais e informações sobre o tratamento de dados pessoais

Cada operação de tratamento de dados pessoais precisa ter um fundamento jurídico, ou seja, uma base legal, que nada mais é que uma justificativa que a autorize, prevista na Lei Geral de Proteção de Dados Pessoais.

Todas as Nossas atividades de tratamento de dados pessoais possuem uma base legal que as fundamenta, dentre as permitidas pela legislação. Mais informações sobre as bases legais que utilizamos para operações de tratamento de dados pessoais específicas podem ser obtidas a partir de nossos canais de contato informados ao final desta Política.

##### 2. Dados que coletamos, onde são coletados e motivos da coleta

Nosso modelo coleta e utiliza alguns dados pessoais dos pacientes do Instituto do Câncer do Estado de São Paulo, conforme o disposto nesta seção.

### *1. Dados sensíveis*

O modelo poderá coletar os seguintes dados sensíveis dos usuários:

- Dados sobre a origem racial ou étnica;
- Dados genéticos;
- Dados relativos à saúde do usuário;
- Dados relativos à vida sexual do usuário.

Eles são coletados por intermédio do ICESP visando fornecer o melhor tratamento possível aos pacientes.

O grupo G4 usará os dados exclusivamente para fins médicos, com o intuito de sugerir o melhor tratamento possível para o bem-estar do paciente. Não compartilhamos esses dados com terceiros sem o consentimento explícito dos pacientes e do ICESP.

Em qualquer caso, o tratamento de dados pessoais sensíveis somente ocorrerá para atender a finalidades específicas expressas nesta política ou devidamente informadas ao usuário por outros meios.

### **3. Por quanto tempo seus dados pessoais serão armazenados e forma de armazenamento**

Os dados pessoais coletados pelo modelo são armazenados e utilizados durante o período que corresponda ao necessário para atingir as finalidades elencadas neste documento e que considere os direitos de seus titulares, conforme o disposto no inciso I do artigo 15 da Lei 13.709/18.

Eles podem ser removidos ou anonimizados a pedido do usuário, excetuando os casos em que a lei oferecer outro tratamento.

Os dados coletados são armazenados em ambiente seguro e em servidor próprio ou de terceiro contratado para este fim.

### **4. Compartilhamento de dados pessoais com terceiros**

Nós compartilhamos alguns dos dados pessoais mencionados nesta seção com terceiros. Os dados coletados serão, após o tratamento, compartilhados com os

médicos responsáveis pelo diagnóstico e tratamento de casos de câncer de mama, que utilizarão o modelo preditivo.

O grupo G4 usará os dados dos pacientes exclusivamente para fins médicos, com o intuito de sugerir ao médico o melhor tratamento possível. Não compartilharemos esses dados com terceiros sem o consentimento explícito dos pacientes e do ICESP.

Além das situações aqui informadas, é possível que compartilhem dados com terceiros para cumprir alguma determinação legal ou regulatória, ou ainda, para cumprir alguma ordem expedida por autoridade pública.

Em qualquer caso o compartilhamento de dados pessoais observará todas as leis e regras aplicáveis, buscando sempre garantir a segurança dos dados de nossos usuários, observados os padrões técnicos empregados no mercado.

## **5. Cookies ou dados de navegação**

Os cookies referem-se a arquivos de texto enviados pela plataforma ao computador do usuário e visitante e que nele ficam armazenados, com informações relacionadas à navegação no site. Tais informações são relacionadas aos dados de acesso como local e horário de acesso, sendo armazenadas pelo navegador do usuário e visitante para que o servidor da plataforma possa lê-las posteriormente a fim de personalizar os serviços da plataforma.

O usuário e o visitante da plataforma manifesta conhecer e aceitar que pode ser utilizado um sistema de coleta de dados de navegação mediante à utilização de cookies.

O cookie persistente permanece no disco rígido do usuário e visitante depois que o navegador é fechado e será usado pelo navegador em visitas subsequentes ao site. Os cookies persistentes podem ser removidos seguindo as instruções do seu navegador. Já o cookie de sessão é temporário e desaparece depois que o navegador é fechado. É possível redefinir seu navegador da web para recusar todos os cookies, porém alguns recursos da plataforma podem não funcionar corretamente se a capacidade de aceitar cookies estiver desabilitada.

## 5. Como o titular pode solicitar e exercer seus direitos

O Titular tem direito a obter do G4 ou ICESP, em relação aos dados por ele tratados, a qualquer momento, e mediante requisição:

- a) confirmação da existência de tratamento;
- b) acesso aos dados;
- c) correção de dados incompletos, inexatos ou desatualizados;
- d) anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto na Lei n.º 13.709, de 2018;
- e) portabilidade dos dados a outro empregador, mediante requisição expressa e observados os dispositivos da lei trabalhista, conforme a regulamentação do órgão Controlador;
- f) eliminação dos dados pessoais tratados com o consentimento do(a) empregado(a), exceto nas hipóteses previstas no Art. 16 da Lei n.º 13.709, de 2018;
- g) informação das entidades públicas e privadas com as quais o Controlador realizou uso compartilhado de dados;
- h) informação sobre a possibilidade de não fornecer consentimento e sobre as consequências da negativa;
- i) revogação do consentimento, nos termos do § 5º do Art. 8º da Lei n.º 13.709, de 2018.

Para garantir que o usuário que pretende exercer seus direitos é, de fato, o titular dos dados pessoais, poderemos solicitar documentos ou outras informações que possam auxiliar em sua correta identificação, a fim de resguardar nossos direitos e os direitos de terceiros. Isto será somente feito, porém, se for absolutamente necessário, e o requerente receberá todas as informações relacionadas.

## 6. Medidas de segurança no tratamento de dados pessoais

Empregamos medidas técnicas e organizativas aptas a proteger os dados pessoais de acessos não autorizados e de situações de destruição, perda, extravio ou alteração desses dados.



As medidas que utilizamos consideram a natureza dos dados, o contexto e a finalidade do tratamento, os riscos que uma eventual violação geraria para os direitos e liberdades do usuário, e os padrões atualmente empregados no mercado por empresas semelhantes ao nosso grupo.

O grupo G4 armazenará os dados dos pacientes de forma segura e protegida, usando medidas de segurança físicas e digitais. Apenas os funcionários autorizados terão acesso aos dados, e o modelo utilizado não estará livre para uso público, assim como os dados nele utilizados.

De qualquer forma, caso ocorra qualquer tipo de incidente de segurança que possa gerar risco ou dano relevante para qualquer de nossos usuários, comunicaremos os afetados e a Autoridade Nacional de Proteção de Dados acerca do ocorrido, conforme o disposto na Lei Geral de Proteção de Dados.

## **7. Alterações nesta política**

A presente versão desta Política de Privacidade foi atualizada pela última vez em: 24/02/2023.

Reservamo-nos o direito de modificar, a qualquer momento, as presentes normas, especialmente para adaptá-las às eventuais alterações feitas em nosso modelo, seja pela disponibilização de novas funcionalidades, seja pela supressão ou modificação daquelas já existentes.

Sempre que houver uma modificação, nossos usuários serão notificados acerca da mudança.

## **8. Como entrar em contato conosco**

Para esclarecer quaisquer dúvidas sobre esta Política de Privacidade ou sobre os dados pessoais que tratamos, entre em contato pelos canais oficiais do ICESP, ou com nosso Encarregado de Proteção de Dados Pessoais, pelo canal mencionado abaixo:

E-mail: [inteli@inteli.edu.br](mailto:inteli@inteli.edu.br)



## 4.3. Preparação dos Dados e Modelagem

Caso seu projeto seja:

1. Modelo supervisionado:

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

Caso seu projeto seja:

1. Modelo não-supervisionado:

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

## 4.4. Comparação de Modelos

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

a) Escolha da métrica e justificativa.

b) Modelos otimizados.

- Apresentar três modelos e suas métricas.

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

c) Definição do modelo escolhido e justificativa.

## 4.5. Avaliação

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

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

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

## 5. Conclusões e Recomendações

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

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

## 6. Referências

*Incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.*

*Um exemplo de referência de livro:*

*LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.*

*SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.*

*<https://www.nngroup.com/articles/journey-mapping-101/> (Referência 4.1.7)*

## Anexos

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