



**OncoAI**  
**Faculdade de**  
**Medicina da USP**



## Controle do Documento

### Histórico de revisões

Data	Autor	Versão	Resumo da atividade
02/02/2023	Mariana B. Görresen	1.1	Preenchimento das seções 4.1.3.1, 4.1.3.4
05/02/2023	Mariana B. Görresen	1.2	Preenchimento da seção 4.1.4
05/02/2023	Bruno Wasserstein	1.3	Preenchimento da seção 2
07/02/2023	Mariana B. Görresen	1.4	Preenchimento da seção 4.1.1.1
08/02/2023	Stefano Parente	1.5	Preenchimento das seções 4.1.2, 4.1.5
08/02/2023	Mauricio Felicissimo	1.6	Preenchimento das seções 4.1.6 e 4.1.7
09/02/2023	Stefano Parente	1.7	Preenchimento das seções 4.1.3.2, 4.1.3.3
09/02/2023	João Montagna	1.8	Preenchimento das seções 1 e 4.1.1.2
10/02/2023	Mariana B. Görresen	1.9	Revisão dos tópicos preenchidos
14/02/2023	Bruno Wasserstein e Mariana B. Görresen	2.0	Correção de detalhes e modificação de textos
15/02/2023	Stefano Parente	2.1	Correção da seção 4.1.2
23/02/2023	Bruno Wasserstein e Stefano Parente	2.2	Preenchimento da seção 4.1.8
23/02/2023	Bruno Wasserstein	2.3	Preenchimento da seção 4.2.3 e letra c) da seção 4.2.1
24/02/2023	Mariana Görresen	2.4	Preenchimento da seção 4.2.2 e letras a) e b) da seção 4.2.1
08/03/2023	Stefano Parente	2.5	Preenchimento da seção 3
08/03/2023	Bruno Wasserstein	2.6	Preenchimento da seção 4.3.1 e adequação das referências bibliográficas para as normas da ABNT
10/03/2023	Bruno Wasserstein e Mariana Görresen	2.6.1	Preenchimento das seções 4.3.2 e 4.3.3

# 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. 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
4.1.8. LGPD	
4.2. Compreensão dos Dados	9
4.2.1. Exploração dos dados	
4.2.2. Pré processamento	
4.2.3. Hipóteses	
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

Fundada em 2008, o Instituto do Câncer do Estado de São Paulo (ICESP) é uma instituição de pesquisa e tratamento médico situada em São Paulo. O Instituto não é somente um dos melhores hospitais da América Latina, também é um dos maiores centros de referência em Oncologia. Possui um grande centro com cerca de 70.000 m<sup>2</sup> e mais de 445 leitos, além de dispor de uma equipe com diversas especialidades médicas em relação ao tratamento de câncer.

O ICESP tem uma enorme relevância para a humanidade, pois o propósito do instituto é ter um foco exclusivo para o tratamento de câncer, oferecendo um cuidado de alta qualidade e dispondo de tecnologias de ponta. Além disso, ele é reconhecido mundialmente pela estrutura de pesquisa, contribuindo para o desenvolvimento de novas terapias.

Atualmente, a escolha entre os tratamentos adjuvante e neoadjuvante, na área do câncer de mama, enfrenta alguns desafios, devido a diversas variáveis como o estágio da doença, tamanho do tumor e outros que podem afetar o sucesso ou não do tratamento. Além disso, as decisões entre esses tratamentos ainda estão sujeitas a erros humanos e imprecisões, como avaliação do patologista e variações na decisão de tratamento entre diferentes oncologistas. Isso acontece porque falta uma ferramenta de análise mais precisa e assertiva para auxiliar na tomada de decisão.

Por isso, o projeto solicitado pelo ICESP, com Faculdade de Medicina da USP, tem como foco a descoberta de um padrão preditivo entre pacientes diagnosticados com câncer, para determinar o tipo de terapia de tratamento mais adequada, neoadjuvante ou adjuvante, para cada caso.

## 2. Objetivos e Justificativa

### 2.1. Objetivos

Atualmente, o ICESP trabalha com pesquisa, diagnóstico e tratamento de diversos tipos de câncer, entre eles, o câncer de mama. O tratamento de câncer de mama varia dependendo de diversos fatores que devem ser considerados para definir qual terá maior empregabilidade, trazendo mais benefícios para o paciente no longo prazo. Atualmente, por ter que considerar diversas variáveis para definição do melhor tratamento e por ser algo que pode ser mensurável, mesmo com uma margem de incerteza, o ICESP visa conseguir melhorar o êxito nos tratamentos de câncer de mama. Definindo, se o mais adequado é o tratamento adjuvante ou neoadjuvante e qual deve ser a terapia para o caso de tumores com detecção do tipo anti-HER2. Com isso, o Inteli (Instituto de tecnologia e liderança) visa auxiliar o ICESP na decisão de tratamento através da criação de um modelo preditivo de tipo de tratamento para câncer de mama.

### 2.2. Proposta de Solução

O modelo preditivo se baseará nos dados que são relevantes para a decisão do tratamento, então primeiramente será necessário filtrar os dados realmente necessários dos dados gerais que compõem o banco de dados primário. Então, fazendo regressões lineares dos dados citados para criar um modelo preditivo mais preciso. O modelo preditivo levará fatores como tipo do tumor, idade, quantidade de filhos biológicos, para determinar qual a melhor forma de tratamento para o paciente, definindo se o melhor seria o adjuvante ou o neoadjuvante, baseando-se em dados.

Dessa forma, será possível auxiliar os mastologistas na tomada de decisão sobre o tratamento, uma vez que o modelo contará com a análise de dados passados para revisar a taxa de sucesso e ajudar o médico a escolher o tratamento mais adequado. Para definir o tratamento, o médico deverá inserir os dados necessários da paciente para definir qual será o tratamento mais adequado.

### 2.3. Justificativa

A solução se baseia em regressões lineares dos dados disponíveis, considerados necessários e importantes para definição do tipo de tratamento. Uma vez que a análise de dados pode identificar padrões, é possível mensurar considerando as "n" variáveis para definir se será terapia adjuvante ou neoadjuvante. Entre os benefícios estão usar todos os dados necessários disponíveis, não passando despercebido, uma vez que a análise pela máquina é precisa, outro benefício é a velocidade de decisão/resposta de definição de tratamento mais adequado. Potencialmente a IA poderá ser aplicada para definição de outros tratamentos do câncer de mama e possivelmente até para outros tipos de doenças.

## 3. Metodologia

- **Entendimento do negócio:**

Inicialmente, o objetivo desta etapa é realizar um estudo e entender completamente os objetivos de negócio e as necessidades do cliente. Isso envolve a identificação dos objetivos do projeto, o que o cliente deseja alcançar, como os resultados serão utilizados (aplicações para o produto) e quais são os possíveis impedimentos e riscos do projeto.

- **Entendimento dos dados:**

Nessa etapa, o objetivo é coleta, exploração e mineração dos dados. Isso inclui a identificação dos dados necessários para alcançar os objetivos do projeto, a obtenção dos dados, a avaliação da qualidade e relevância dos dados e determinação se os dados são suficientes para atender o propósito ou se precisam passar por transformação ou limpeza.

- **Preparação dos dados:**

Nesta fase, antes de iniciar a construção da solução, os dados são preparados, transformados e adaptados para a análise. Isso envolve a seleção dos dados de maior relevância, o tratamento de valores ausentes ou duplicados, limpeza destes dados e a seleção de características que serão utilizadas na análise.

- **Modelagem:**

O objetivo desta fase, a partir da preparação dos dados e sua divisão entre treino e teste, é criar um modelo estatístico que possa ser usado para prever ou classificar dados. Isso envolve: determinar os algoritmos a serem testados, a definição dos parâmetros, a avaliação do desempenho e a validação do modelo, **com interpretação dos resultados.**

- **Avaliação:**

Nesta fase, o modelo desenvolvido é avaliado em relação aos critérios de sucesso do negócio. Isso envolve a avaliação do desempenho do modelo em dados de teste e a avaliação em termos de negócio. Aqui, serão definidos os próximos passos, envolvendo as possíveis ações e decisões a serem tomadas.

- **Implantação:**

Na etapa final, o objetivo é implantar o modelo criado em um ambiente de produção. Isso envolve a integração do modelo, a configuração do ambiente de produção e deve ser elaborado um relatório final apresentando os resultados obtidos e alternativas de ação.

## 4. Desenvolvimento e Resultados

### 4.1. Compreensão do Problema

#### 4.1.1. Contexto da indústria

##### 4.1.1.1. Análise da indústria

Ao analisar o contexto da do ICESP (Instituto do Câncer do Estado de São Paulo) como instituição pública. Percebe-se que seu objetivo é auxiliar a população da sociedade brasileira com tratamentos, pesquisas e desenvolvimento tecnológico para a medicina. Podemos identificar possíveis concorrentes nesta área, focando, principalmente, no âmbito de tratamento e pesquisa do câncer de mama. Estes são:

- O Centro de Oncologia e Hematologia Einstein Família Dayan – Daycoval, do Hospital Albert Einstein, que ficou em primeiro lugar entre os melhores hospitais da América Latina e vigésimo primeiro em melhores hospitais oncológicos do mundo, em 2021.
- O Instituto Nacional de Câncer (INCA), que, ao contrário do Hospital Albert Einstein, citado acima, é um instituto e hospital vinculado ao governo e possui o mesmo público e missões do parceiro do projeto. O instituto promove a assistência, prevenção, ensino e pesquisa, ou seja, atua com o mesmo objetivo do ICESP.
- O Hospital A.C. Camargo Cancer Center, é um hospital oncológico localizado na cidade de São Paulo. Tem foco em diagnóstico, tratamento e pesquisa sobre câncer.

O ICESP é uma instituição pública, sendo assim, sem fins lucrativos, com foco em ser referência de tratamento, pesquisa e desenvolvimento de inovações para a indústria da medicina. Seu modelo de negócios se baseia na pesquisa e tratamento de câncer, havendo gerenciamento público do Estado de São Paulo, incluindo investimentos governamentais, além de doações privadas.

A medicina tem evoluído constantemente para fornecer tratamentos mais eficazes e atendimento de qualidade aos pacientes. Uma das tendências atuais é a personalização do tratamento, onde a Inteligência Artificial é utilizada para desenvolver modelos preditivos que consideram as características únicas de cada paciente. Isso permite uma análise mais precisa dos dados, resultando em tratamentos mais eficazes e seguros, aumentando, assim, as chances de sucesso no tratamento. No câncer de mama, por exemplo, o Machine Learning também é uma ferramenta que pode permitir prever, com a utilização de algoritmos genéticos, a tendência do desenvolvimento, de risco e o tipo do câncer muito antes de se tornar perigoso para o paciente. Desta forma, com a ajuda das tendências indicadas acima, os médicos podem tomar decisões mais assertivas sobre o tratamento ideal para cada paciente, tanto no ramo médico em geral quanto, principalmente, no foco do nosso projeto, o câncer de mama.



#### 4.1.1.2. As 5 Forças de Porter

**Rivalidade entre os concorrentes:** Pode haver concorrência indireta de outras instituições médicas que oferecem tratamentos relacionados ao câncer, porém o instituto tem um nome renomado especializado no tratamento de câncer, assim obtendo preferência em geral e saindo na frente de seus possíveis concorrentes.

**Ameaça de novos entrantes:** O ICESP possui uma reputação consolidada que pode influenciar para o surgimento de novos entrantes nesse mercado. Porém, a entrada de novos concorrentes pode ser considerada relativamente fácil para instituições de saúde, especialmente se houver uma demanda por serviços de qualidade.

**Poder de barganha dos fornecedores:** O poder de barganha dos fornecedores como laboratórios de exames ou indústria farmacêutica não chega a ser tão alto, uma vez que os mesmos fornecem produtos que não são tão raros no mercado. Porém, o poder de barganha de fornecedores de equipamentos médicos e hospitalares são de alta influência, já que é uma indústria altamente monopolizada.

**Poder de barganha dos compradores:** O poder de barganha dos compradores não pode ser considerado alto, visto que eles são pacientes. Isso significa, que estão em uma situação de baixo poder de negociação de tratamentos e seus preços, mas, alta necessidade dos serviços da instituição.

**Ameaça de produtos substitutos:** A ameaça de produtos substitutos é mediana. Por exemplo, clínicas particulares e hospitais que oferecem tratamento de câncer são considerados ameaças, além de outros institutos de pesquisa com alto investimento. Porém, o ICESP possui recursos extremamente avançados, onde tem serviços de tratamentos únicos, tornando-os difíceis de serem substituídos.

## 4.1.2. Análise SWOT

ICESP

### Análise SWOT

#### Strengths

- O ICESSP é uma instituição reconhecida como referência no tratamento do câncer no Brasil e em outras partes do mundo;
- O instituto conta com equipamentos de alta tecnologia, o que possibilita o diagnóstico e tratamento mais precisos e eficazes;
- Atendimento humanizado e personalizado, o que melhora a qualidade de vida durante o tratamento;

#### Weaknesses

- O número de pacientes em busca de tratamento no ICESSP vem aumentando constantemente, o que pode sobrecarregar a instituição e prejudicar o atendimento;
- O instituto está localizado em São Paulo, o que pode dificultar o acesso de pacientes de outras regiões do país;
- O ICESSP é uma instituição pública e, portanto, depende de recursos públicos para operar;

#### Opportunities

- Ampliar parcerias estratégicas com outras instituições, para gerar mais oportunidades de atuação e oferta de serviços;
- A conscientização sobre o câncer vem aumentando no Brasil, o que pode gerar mais demanda por serviços do ICESSP;
- O ICESSP é uma instituição que aceita doações, e um aumento dessas doações pode gerar mais investimentos em equipamentos e tecnologias;

#### Threats

- O ICESSP enfrenta a concorrência de outras instituições de saúde no tratamento do câncer;
- O surgimento de novas tecnologias e tratamentos pode tornar obsoletas as técnicas e equipamentos utilizados pelo ICESSP, o que exige investimentos constantes em atualizações e aquisição de novas tecnologias;
- Mudanças no sistema de saúde podem afetar o financiamento e a capacidade de atuação do ICESSP;

## 4.1.3. Planejamento Geral da Solução

### 4.1.3.1. Qual é o problema a ser resolvido:

Os médicos de câncer de mama estão tendo problemas na escolha do tratamento conforme o tipo de câncer de mama do paciente apresentado. Esse problema infere também no fato de que há uma grande dificuldade na análise de dados disponíveis no seu banco de dados e falta de ferramenta auxiliar para a melhor escolha de decisão.

#### **4.1.3.2. Qual a solução proposta (visão de negócios):**

Desenvolvimento de um modelo preditivo para aconselhamento da melhor opção de tratamento para o câncer de mama, baseado em uma análise rigorosa dos dados previamente coletados de pacientes. Esse modelo utiliza técnicas avançadas de análise de dados e aprendizado de máquina para prever qual será o melhor tratamento para cada paciente individualmente, considerando os fatores como histórico médico, estágio da doença e resposta a tratamentos anteriores. A finalidade é fornecer aos profissionais de saúde informações precisas e confiáveis para ajudar na decisão de tratamento e garantir a melhor qualidade de vida para as pacientes com câncer de mama.

#### **4.1.3.3. Como a solução proposta deverá ser utilizada:**

A solução proposta deverá ser utilizada por médicos mastologistas para auxiliar na escolha do tratamento entre os disponíveis para o câncer de mama, onde através da resposta dada pela IA com dados essenciais e tratados sobre qual tratamento indicado, o médico analisará e dará o diagnóstico final ao paciente. A intenção é proporcionar uma ferramenta para os médicos poderem tomar decisões mais informadas e assertivas quanto ao tratamento do câncer.

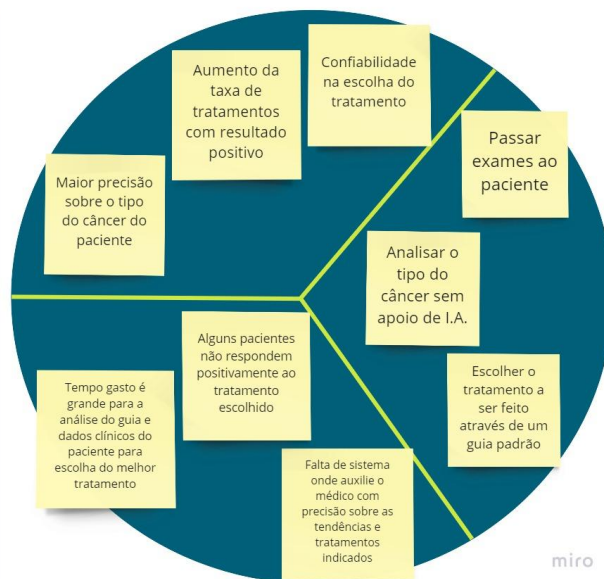
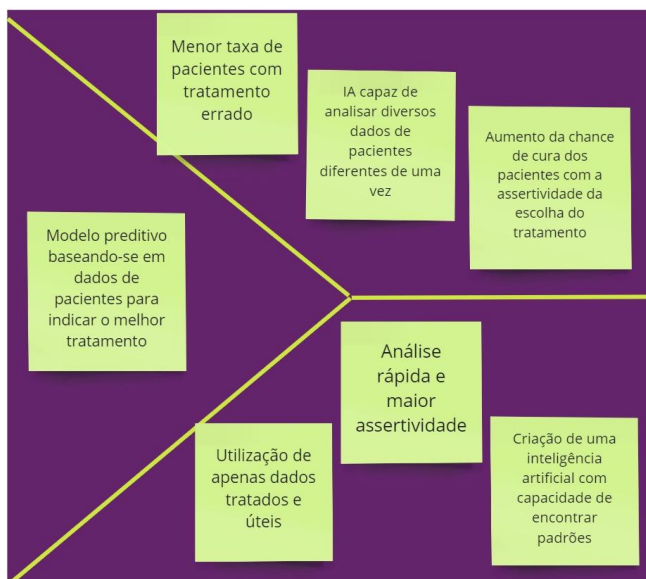
#### **4.1.3.4. Quais os benefícios trazidos pela solução proposta:**

Os principais benefícios são o aumento da posição do médico no diagnóstico e escolha do tratamento, pelo auxílio dos resultados vindos da IA. Através desse auxílio, o médico terá maior chance de ter indicado o melhor tratamento para aquele caso. Sendo assim, aumentará a taxa de pacientes com possível cura do câncer de mama.

#### **4.1.3.5. Qual será o critério de sucesso e qual medida será utilizada para o avaliar:**

O critério de sucesso será o treinamento da máquina a partir de dados anteriores, para dessa forma identificar padrões de sucesso de tratamento, para quando forem inseridos os dados omitidos da máquina, ela indique o tratamento correto, uma vez que esses dados já têm os tratamentos prescritos corretamente. A medida para avaliar será baseada em taxas de sucesso de acertos aceitáveis para modelos preditivos, considerando fator como a quantidade de dados, experiência dos desenvolvedores, para criar uma taxa de sucesso aceitável.

## 4.1.4. Value Proposition Canvas



## 4.1.5. Matriz de Riscos

Matriz de Risco						
Probabilidade		Riscos				
Muito Alta	5					O profissional de saúde/paciente não conseguir usar o programa adequadamente
Alta	4				Mudanças na medicina, resultando em modelos desatualizados e imprecisos	
Médio	3		Algum membro do grupo não contribuir		Modelo pode se ajustar demais aos dados de treinamento, resultando em previsões ineficazes para novos pacientes	
Baixa	2			Impacto ético: uso inadequado ou impacto negativo pode ter implicações éticas e sociais		Erro de cálculo que resulte em uma escolha de tratamento imprecisa
Muito Baixa	1	Erro ortográfico na resposta de recomendação do tratamento		Falta de confiança do paciente		Vazamento de dados pessoais e clínicos
		1	2	3	4	5
		Muito Baixo	Baixo	Médio	Alta	Muito Alta
						Impacto

Oportunidade				
Indicar com maior precisão qual o melhor tratamento para cada paciente		Redução de custos (relacionados ao tratamento e monitoramento da doença)		
Melhora na velocidade da decisão clínica	Estudo de fatores de risco e prevenção do câncer de mama	Monitoramento da resposta ao tratamento		
	Descobrir padrões que influenciam na eficácia do tratamento	Melhor comunicação entre profissionais de saúde, tendo uma abordagem mais coordenada e eficaz		
				Modelo preditivo ser adotado como revisor de análise padrão
5	4	3	2	1
Muito Alta	Alta	Médio	Baixo	Muito Baixo
to				

## 4.1.6. Personas

Persona 1 (utiliza o modelo):

### PERFIL

Nome : Maria Beatriz  
Idade : 40  
Ocupação : Mastologista  
Educação : Ensino Superior

### BIOGRAFIA

Dra. Maria Beatriz, mastologista com 14 anos de experiência em diagnóstico e tratamento de câncer de mama. Ela tem como principal objetivo fornecer o melhor cuidado para seus pacientes. A Dra. está sempre em busca de novas tecnologias e soluções que possam ajudá-la a aprimorar o diagnóstico e o tratamento do câncer de mama. Ela acredita que um modelo preditivo de câncer de mama seria uma ferramenta valiosa para ajudá-la na tomada de decisões clínicas informadas e aumentar a precisão do diagnóstico, melhorando a qualidade de vida de seus pacientes.

### PERSONALIDADE

Comunicativa

Analítica

Ansiosa

Divertida

Otimista

Bem-humorada

Independente



### INTERESSES

- Avanços em tratamento de mama
- Biologia
- Acompanhamento de pacientes
- Estudo de casos complexos

### INFLUÊNCIAS

- Avanços científicos e médicos
- Outros especialistas da área
- Experiência clínica
- Faculdade de Medicina da USP
- Feedback de colegas

### METAS

- Ajudar pacientes com câncer de mama
- Desenvolvimento de técnicas mais eficientes
- Desenvolvimento de novos tratamentos
- Reputação profissional

### NECESSIDADES E EXPECTATIVAS

- Mais tempo para avaliação e tratamento
- Prestar diagnósticos mais precisos
- Confirmar o prognóstico oferecido

### MOTIVAÇÕES

- Satisfação em ver pacientes melhorarem
- Aumentar a confiança na saúde pública

### DORES E FRUSTRAÇÕES

- Casos avançados ou inoperáveis
- Falta de recursos
- Longas horas de trabalho
- Dificuldade em oferecer diagnóstico impreciso

Persona 2 (utiliza o modelo):

### PERFIL

Nome : Jessica Almeida  
Idade : 45  
Ocupação : Vendedora  
Educação : Ensino Medio



### BIOGRAFIA

Jessica Almeida é uma mulher de 45 anos, natural de São Paulo, e trabalha como vendedora em uma loja de departamentos. Ela é casada há 19 anos com seu marido, com quem tem dois filhos adultos, e luta contra a obesidade e o estilo de vida sedentário. Infelizmente, a vida de Jessica mudou drasticamente quando ela foi diagnosticada com câncer de mama. Apesar de ser uma mulher forte e corajosa, o diagnóstico foi uma notícia chocante e assustadora para ela. Mas, apesar das dificuldades, Jessica está determinada a vencer o câncer.

### PERSONALIDADE

Introvertida ☐

Engraçada ☐

Ansiosa ☐

Compulsiva ☐

Sedentária ☐

Insegura ☐

Corajosa ☐

### INTERESSES

- Passar tempo com a família e amigos
- Cuidar do jardim
- Ler romances e assistir a filmes de drama
- Culinária
-

### INFLUÊNCIAS

- Família e amigos próximos
- Grupos de apoio a pacientes com câncer de mama
- Mídia social e artigos de saúde
- Livros e filmes sobre superação e perseverança

### METAS

- Seguir o tratamento médico com determinação
- Perder peso e seguir uma dieta saudável
- Aprender a lidar com o estresse e a ansiedade
- Continuar trabalhando e cuidando de sua casa

### NECESSIDADES E EXPECTATIVAS

- Ter acesso a equipe médica disponível e atenciosa
- Receber apoio emocional da família e amigos
- Ter acesso a recursos financeiros para cobrir as despesas médicas

### MOTIVAÇÕES

- Proteger sua família e continuar a cuidar deles
- Melhorar sua qualidade de vida e bem-estar
- Ser uma inspiração para outras pessoas com câncer

### DORES E FRUSTRAÇÕES

- Medo de não vencer o câncer
- Sentir-se sozinha e incapaz de lidar com a doença
- Preocupação com as despesas médicas e financeiras



## 4.1.7. Jornadas do Usuário

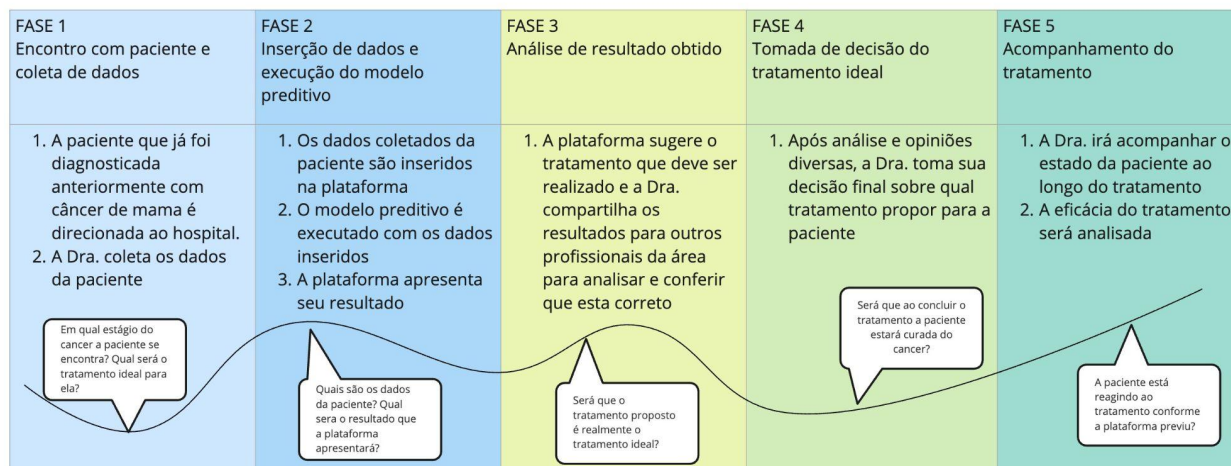


### Dra. Maria Beatriz

**Cenário:** A Dra. Maria Beatriz requer ajuda para tomar a decisão de qual é o tratamento de câncer de mama ideal para seus pacientes. Ela precisa da ajuda de um modelo preditivo para tomar a decisão certa com mais eficiência e assertividade.

### Expectativas

- Uma plataforma intuitiva e fácil de usar
- Resultados mais precisos
- Realizar um diagnóstico de forma mais eficiente



### Oportunidades

Melhorar e facilitar a usabilidade da plataforma, deixando-a mais intuitiva. Desenvolver uma função onde após apresentar o resultado, a plataforma expõe um relatório explicando como chegou ao resultado e por que o tratamento proposto é o ideal.

### Responsabilidades

A plataforma apresenta um resultado com base aos dados que foram inseridos, então cabe à equipe médica certificar-se que tais dados estão corretos. Também, por mais que a plataforma apresenta resultados com mais assertividade, os médicos são responsáveis por analisar os resultados e conferir que o tratamento proposto seja de fato o tratamento ideal.

miro

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

### Política de Privacidade da OncoAI

A OncoAI é uma equipe dedicada ao desenvolvimento de modelos preditivos de tratamento de câncer de mama, comprometida em proteger a privacidade dos dados pessoais fornecidos pelos usuários. A presente Política de Privacidade tem como objetivo esclarecer aos usuários como a OncoAI coleta, utiliza, armazena e compartilha esses dados, em conformidade com a Lei Geral de Proteção de Dados (LGPD) do Brasil.

#### Coleta de dados:

Os dados utilizados para treinar e validar o modelo são fornecidos pela Faculdade de Medicina da Universidade de São Paulo (USP) de forma anônima e agregada, sem identificação individual dos pacientes. Esses dados incluem informações clínicas e de exames, coletados de fontes públicas e privadas.

Além disso, a OncoAI pode coletar informações não identificáveis, como endereço IP, localização geográfica, dados de navegação, sistema operacional, entre outros, por meio do uso de cookies e tecnologias semelhantes.



#### Uso dos dados:

Os dados coletados são usados exclusivamente para fins de pesquisa e desenvolvimento do modelo preditivo de tratamento de câncer de mama. O modelo e os dados associados só serão compartilhados com profissionais médicos autorizados e conforme as leis de privacidade e proteção de dados aplicáveis.

#### Armazenamento de dados:

Os dados coletados pela OncoAI ficarão armazenados em servidores protegidos por medidas de segurança rigorosas, para proteger os dados coletados contra acesso não autorizado, uso inadequado, alteração ou destruição. Os dados serão mantidos enquanto forem necessários para a finalidade para a qual foram coletados ou até que o titular solicite a exclusão.

#### Compartilhamento de dados:

A OncoAI não compartilhará os dados pessoais dos usuários com terceiros sem o consentimento explícito dos mesmos. No entanto, pode haver o compartilhamento de informações agregadas e anônimas para fins de pesquisa e desenvolvimento de novos modelos, desde que isso não comprometa a privacidade dos usuários.

#### Segurança de dados:

A OncoAI adota medidas de segurança técnicas e organizacionais adequadas para proteger os dados coletados contra acesso não autorizado, uso inadequado, alteração ou destruição.

#### Direitos dos usuários:

O titular dos dados tem o direito de acessar, corrigir ou excluir seus dados pessoais a qualquer momento, bem como revogar seu consentimento para o uso desses dados. A OncoAI fornecerá todas as informações e ferramentas necessárias para os usuários poderem exercer seus direitos eficientemente.

#### Atualização da política de privacidade:

Esta política pode ser atualizada periodicamente para refletir mudanças em nossos processos ou em leis aplicáveis. Qualquer atualização será publicada em nossa documentação.

## 4.2. Compreensão dos Dados

### 4.2.1. Exploração de dados:

a) Cite quais são as colunas numéricas e categóricas:

Primeiro, para a padronização de colunas de data/tempo, através da biblioteca pandas, utilizamos o método `'to_datetime'` para padronizar o tipo do dado como `'datetime64'`.

Código:

```
df3['data_entrada'] = pd.to_datetime(df3['data_entrada'])
```

Para a identificação das colunas numéricas, foi utilizado o método `'select_dtypes'`, incluindo apenas valores numéricos, e assim, listando as colunas ditas como numéricas.

Código:

```
numeric_cols = merge_df.select_dtypes(include=np.number).columns.tolist()
numeric_cols
```

Output (nome das colunas numéricas):

```
['record_id',
 'idade_no_primeiro_diagnostico',
 'peso_inicial',
 'altura_cm',
 'IMC',
 'peso_max',
 'peso_min',
 'repeat_instance_x',
 'grau_histopatologico',
 'subtipo_tumoral',
 'indice_h_receptor_de_progesterona',
 'ki67_percentage',
 'repeat_instance_y',
 'codigo_morfologia_de_acordo_com_o_cid_o']
```

Para a identificação das colunas categóricas, foi utilizado o mesmo método, `'select_dtypes'`, incluindo apenas valores considerados `'object'`, e assim, listando as colunas consideradas categóricas.

Código:

```
categorical_cols= merge_df.select_dtypes(include='object').columns.tolist()
categorical_cols
```

Output (nome das colunas categóricas):

```
['sexo',
 'ultima_informacao',
 'ja_ficou_gravida',
```

```

'ja_usou_drogas',
'realiza_atividades_fisicas',
'consumo_de_tabaco',
'consumo_de_alcool',
'possui_historico_familiar_de_cancer',
'grau_de_parentesco(choice=primeiro(pais,_irmaos,_filhos))',
'grau_de_parentesco(choice=segundo(avós,_tios_e_netos))',
'grau_de_parentesco(choice=terceiro(bisavós,tio_avós,primos,sobrinhos))',
'regime_tratamento',
'tipo_de_terapia_anti-her2_neoadjuvante',
'radioterapia',
'esquema_de_hormonioterapia',
'data_entrada',
'diagnostico_primario_tipo_histologico',
'receptor_de_estrogenio',
'receptor_de_progesterona',
'ki67_maior_14_percentage',
'receptor_de_progesterona_quantificacao_percentage',
'receptor_de_estrogenio_quantificacao_percentage',
'her2_por_ihc',
'her2_por_fish',
'data_primeira_consulta_institucional',
'codigo_topografia_cid_0',
'estadio_clinico',
'grupo_estadio_clinico',
'classificacao_tnm_clinico_t',
'classificacao_tnm_clinico_n',
'classificacao_tnm_clinico_m',
'metastase_ao_diagnostico_cid_0_1',
'metastase_ao_diagnostico_cid_0_2',
'metastase_ao_diagnostico_cid_0_3',
'metastase_ao_diagnostico_cid_0_4',
'combinacao_dos_tratamentos_realizados_no_hospital',
'data_recidiva',
'local_recidiva_a_xa0_distancia_metastase_1_cid_0_topografia',
'local_recidiva_a_xa0_distancia_metastase_2_cid_0_topografia',
'local_recidiva_a_xa0_distancia_metastase_3_cid_0_topografia',
'local_recidiva_a_xa0_distancia_metastase_4_cid_0_topografia',
'descricao_da_morfologia_de_acordo_com_o_cid_0_cid_0_3_edição',
'classificacao_tnm_patologico_n',
'classificacao_tnm_patologico_t',
'com_recidiva_a_distancia',
'com_recidiva_regional',
'com_recidiva_local']

```

## b) Estatística descritiva das colunas.

A estatística descritiva foi feita primeiramente para exploração dos dados sem tratamento, assim podendo analisar o estado de cada coluna e identificando possíveis outliers.

Método utilizado para fazer a análise descritiva das colunas numéricas:

```
for i in numeric_cols_para_estatistica:  
    print(df_original[i].describe(), '\n')
```

Exemplo do output:

```
count      3628.000000  
mean        53.933848  
std         13.385260  
min         22.000000  
25%         44.000000  
50%         53.000000  
75%         63.000000  
max         98.000000  
Name: idade_no_primeiro_diagnostico, dtype: float64
```

```
count      45178.000000  
mean        71.237403  
std         241.738021  
min          1.000000  
25%         59.650000  
50%         68.350000  
75%         78.600000  
max        51350.000000  
Name: Peso, dtype: float64
```

```
count      3043.000000  
mean        36.191916  
std         24.598110  
min          0.000000  
25%         16.000000  
50%         30.000000  
75%         50.000000  
max        100.000000  
Name: ki67_percentage, dtype: float64
```

Método utilizado para fazer a análise descritiva das colunas categóricas:

```
for j in categorical_cols_para_estatistica:
    print(df_original[j].value_counts(), '\n')
```

Exemplo do output:

```
Feminino      3627
Masculino      33
Name: sexo, dtype: int64
```

```
Terapia Adjuvante      1294
Terapia Neoadjuvante    1194
Paliativo               57
Não fez quimioterapia    25
Name: regime_tratamento, dtype: int64
```

```
0 (negativo)      2318
+++ (positivo)    1082
++ (duvidoso)     234
+ (negativo)       99
indeterminado      19
Name: her2_por_ihc, dtype: int64
```

### c) Gráficos de correlação da estatística descritiva

Gráfico número 1:

**Gráfico Idade no primeiro diagnóstico X Atividade física**

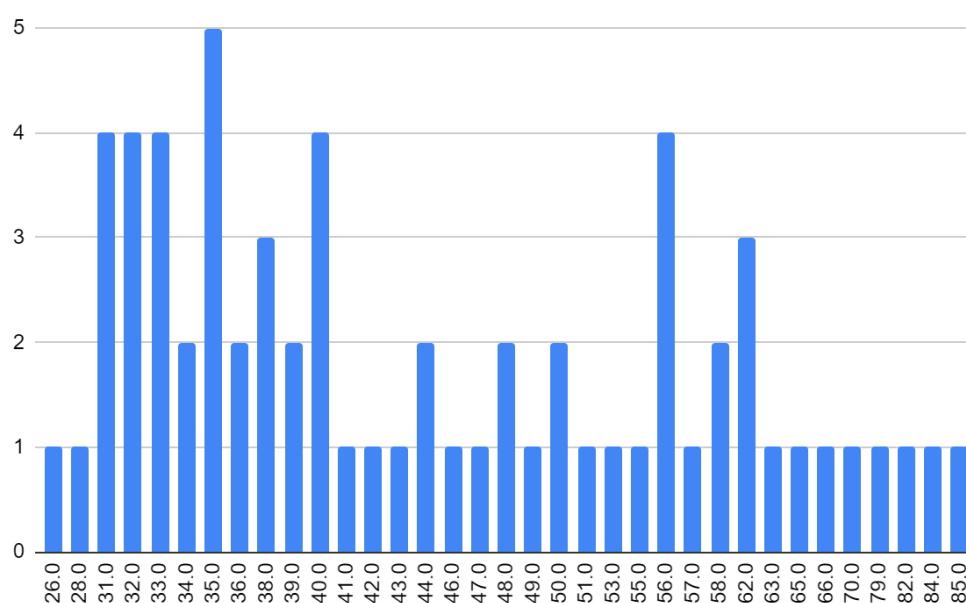
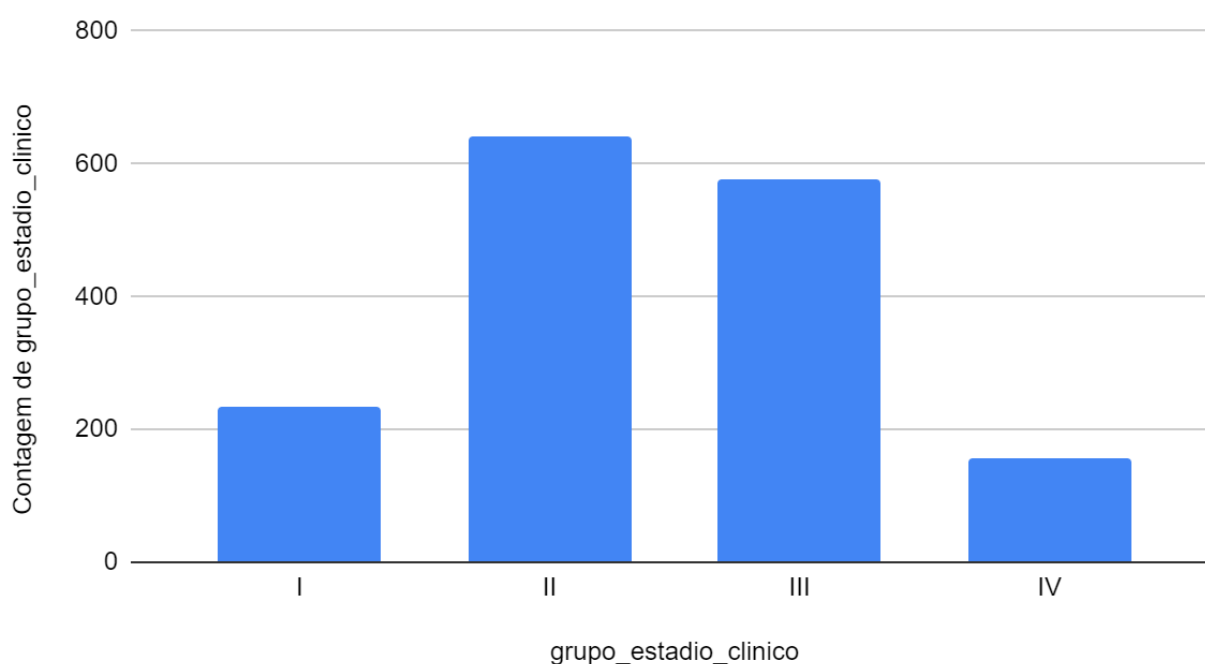


Gráfico que demonstra a relação entre idade e a prática de atividade física (Regular ou frequente) com a idade das pacientes no primeiro diagnóstico, sendo a idade representada no eixo X e a quantidade pacientes que realizam tais atividades no eixo Y.

Conclusão: Através deste gráfico podemos notar que pacientes mais jovens têm uma maior tendência a praticar exercícios com uma regularidade ou frequência, talvez por questões hormonais e físicas. Dessa forma vale observar futuramente a relação entre prática de atividades físicas e sobrevivência e cura do tumor em pacientes que praticam atividade física, uma vez que pesquisas mostram que a prática de atividade física aumenta as chances de tratamento bem-sucedido.

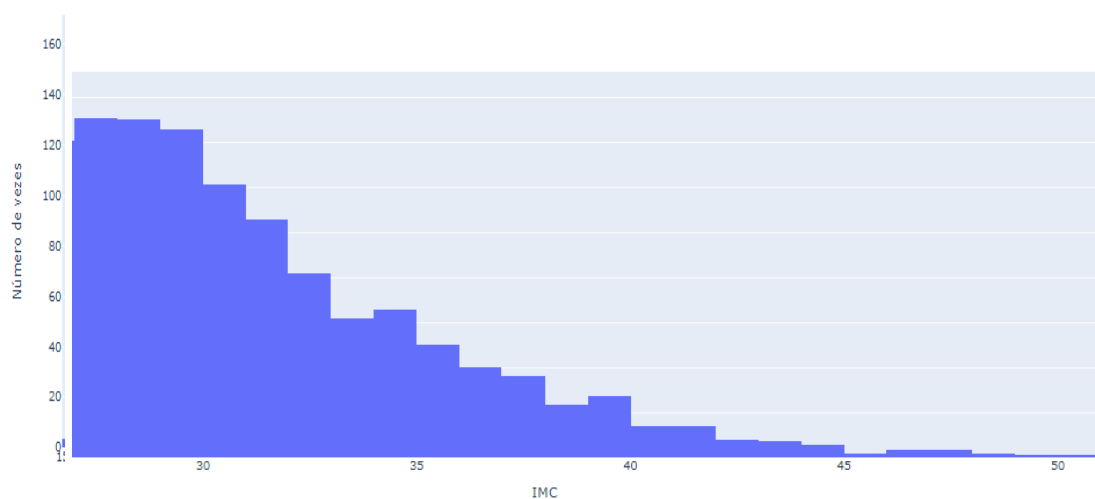
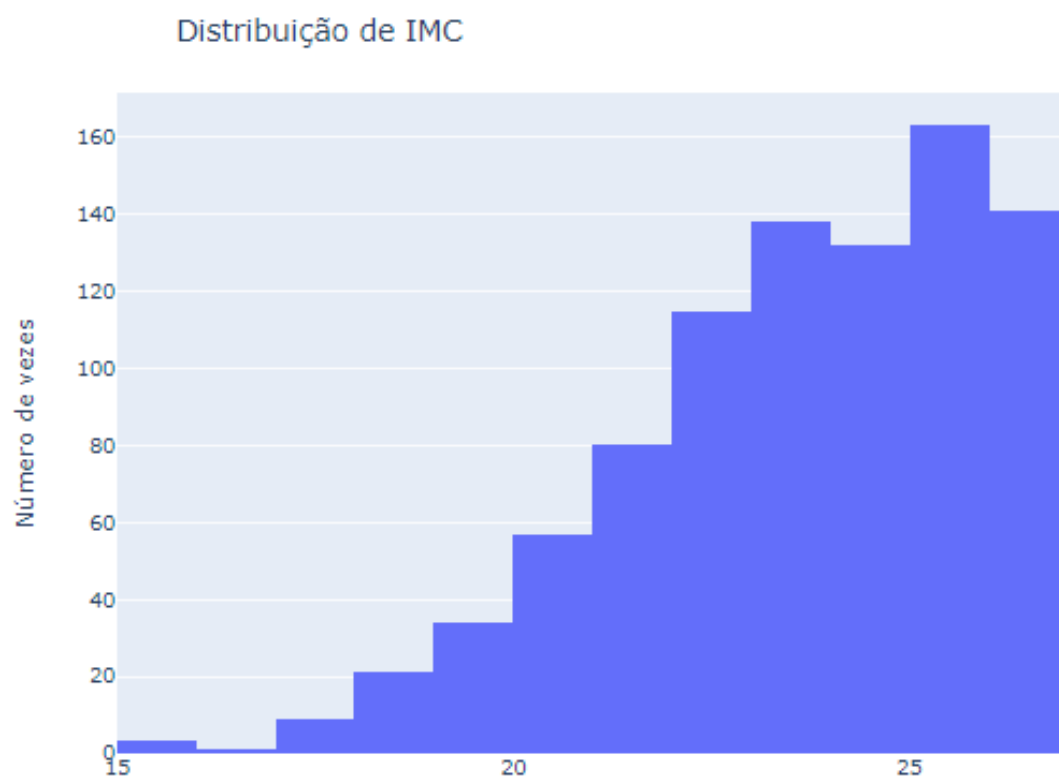
Gráfico número 2:

Contagem de grupo\_estadio\_clinico



Conclusão: Segundo o gráfico número 2 percebe-se a quantidade de pacientes em cada grupo de estágio clínico, é possível notar uma maior quantidade de pacientes nos grupos II e III, demonstrando que os casos intermediários são os que mais ocorrem e casos mais graves têm uma frequência menor que casos mais leves.

Gráfico número 3:



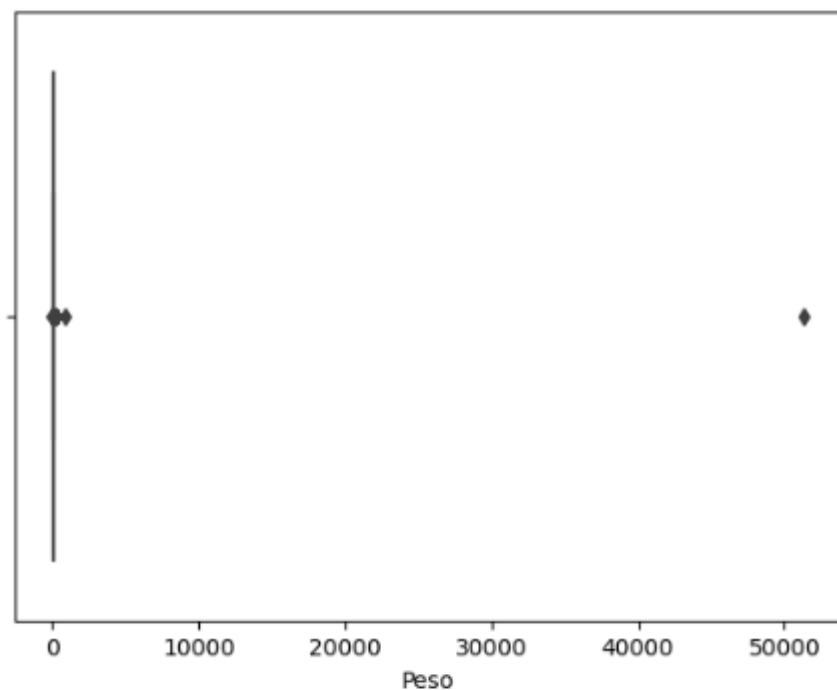
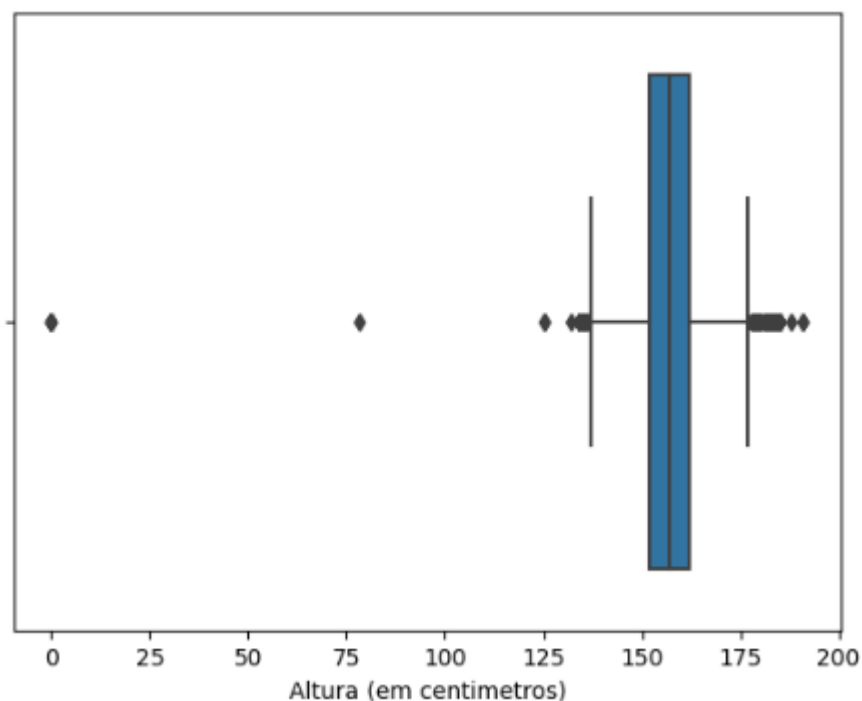




Gráfico número 7:



Nos gráficos 5, 6 e 7, estilo boxplot, pode-se identificar a análise descritiva de cada coluna especificada, essas foram: *idade\_no\_primeiro\_diagnostico*, *Peso* e *Altura (em centímetros)*. Podemos observar os principais pontos de uma coluna para começar a limpeza dos dados, esses são: mediana, q1, q3, limite mais baixo e limite mais alto dos valores, inclusive é possível visualizar possíveis outliers que existem na coluna.

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

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

Visualizamos no processo de tratamento dos dados que haviam outliers nas colunas *Peso*, *idade\_no\_primeiro\_diagnostico* e *Altura (em centímetros)* como foi possível visualizar nos gráficos 5, 6 e 7, na fase de exploração dos dados. Sendo assim, quando começamos a realizar a limpeza, foi investigado como estavam se comportando esses dados nas colunas. Seguem as figuras com os métodos utilizados:

### Verificando os pesos anormais

```
[ ] df3[df3['Peso'] < 30]
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	altura_cm	IMC
23372	43696	Dados Antropometricos	3.0	2014-06-07	29.8	145.75	14.190476
47508	74299	Dados Antropometricos	6.0	2018-01-24	1.0	147.00	0.454545

```
df3[df3['Peso'] > 150]
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	altura_cm	IMC
4085	12651	Dados Antropometricos	3.0	2010-09-14	879.00	159.0	351.600000
9772	21015	Dados Antropometricos	28.0	2011-10-31	51350.00	155.0	21395.833333
15556	27674	Dados Antropometricos	8.0	2017-01-20	154.40	175.0	49.806452
15557	27674	Dados Antropometricos	13.0	2017-12-08	154.40	175.0	49.806452
16161	28542	Dados Antropometricos	21.0	2013-02-17	152.80	178.0	47.750000
16162	28542	Dados Antropometricos	8.0	2013-02-25	153.00	178.0	47.812500
16163	28542	Dados Antropometricos	14.0	2013-05-17	153.80	178.0	48.062500
16164	28542	Dados Antropometricos	4.0	2013-06-10	152.30	178.0	47.593750
16165	28542	Dados Antropometricos	13.0	2013-06-16	151.00	178.0	47.187500
25683	50752	Dados Antropometricos	18.0	2019-02-22	158.00	143.0	79.000000
31196	56902	Dados Antropometricos	24.0	2019-02-06	155.15	157.0	62.060000
43513	70534	Dados Antropometricos	2.0	2017-09-05	177.20	153.0	77.043478

Tivemos como conclusão que os valores 1.00, 879.00 e 51350.00 são claramente outliers providos, provavelmente, de erros de digitação. E então, passamos para a fase de verificar valores levemente acima dos outros, agrupando pelo seu ID e comparando-os, que poderiam, ou não, ser considerados outliers.

df3[df3['Record ID'] == 70534]

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	altura_cm	IMC
43510	70534	Dados Antropometricos	4.0	2017-06-10	114.45	153.0	49.760870
43511	70534	Dados Antropometricos	8.0	2017-07-04	113.20	153.0	49.217391
43512	70534	Dados Antropometricos	1.0	2017-08-15	114.00	153.0	49.565217
43513	70534	Dados Antropometricos	2.0	2017-09-05	177.20	153.0	77.043478
43514	70534	Dados Antropometricos	5.0	2017-10-24	114.80	153.0	49.913043
43515	70534	Dados Antropometricos	7.0	2018-11-28	117.40	153.0	51.043478
43516	70534	Dados Antropometricos	3.0	2019-06-18	117.60	153.0	51.130435
43517	70534	Dados Antropometricos	6.0	2019-07-02	119.20	153.0	51.826087

[ ] df3[df3['Record ID'] == 50752]

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	altura_cm	IMC
25671	50752	Dados Antropometricos	1.0	2014-11-08	50.00	143.0	25.000
25672	50752	Dados Antropometricos	17.0	2017-09-03	48.80	143.0	24.400
25673	50752	Dados Antropometricos	2.0	2018-03-25	46.35	143.0	23.175
25674	50752	Dados Antropometricos	4.0	2018-05-07	46.80	143.0	23.400
25675	50752	Dados Antropometricos	16.0	2018-06-11	47.10	143.0	23.550
25676	50752	Dados Antropometricos	19.0	2018-06-30	47.30	143.0	23.650
25677	50752	Dados Antropometricos	23.0	2018-08-13	46.40	143.0	23.200
25678	50752	Dados Antropometricos	3.0	2018-10-07	47.60	143.0	23.800
25679	50752	Dados Antropometricos	5.0	2018-10-28	47.55	143.0	23.775
25680	50752	Dados Antropometricos	7.0	2018-12-10	48.15	143.0	24.075
25681	50752	Dados Antropometricos	26.0	2018-12-31	47.00	143.0	23.500
25682	50752	Dados Antropometricos	27.0	2019-02-11	45.45	143.0	22.725
25683	50752	Dados Antropometricos	18.0	2019-02-22	158.00	143.0	79.000
25684	50752	Dados Antropometricos	20.0	2019-03-11	45.40	143.0	22.700

Pode-se concluir, com as tabelas acima, que os valores 158.00 e 177.20 são valores incorretos no seu grupo, pois a margem de diferença está muito grande em relação aos outros valores, por isso serão considerados outliers.

Esse método foi utilizado nas colunas *Peso* e *idade\_no\_primeiro\_diagnostico*. A coluna *Altura* (em centímetros) teve seus outliers tratados juntamente com o preenchimento de missings,

onde foi criada uma nova coluna preenchendo novamente as alturas, pegando a mediana do Record ID, assim impedindo que fossem mantidos valores destoantes no mesmo paciente.

```
# Cria uma series para preencher a altura, através do cálculo da mediana, por 'Record ID '
df3_novo= df3.groupby('Record ID')['Altura (em centimetros)'].median()
df3_novo
```

Record ID	Altura (em centimetros)
302	158.0
710	155.0
752	152.0
1367	143.0
1589	167.0
...	
82123	153.0
82124	151.0
82131	156.0
82205	174.0
82240	161.0

Name: Altura (em centimetros), Length: 3803, dtype: float64

```
[70] # Junta a series ao dataframe como coluna em comum 'Record ID', criando uma coluna nova de altura padronizada
df3 = pd.merge(df3, df3_novo, on='Record ID').drop(['Altura (em centimetros)_x'], axis=1)
df3 = df3.rename(columns={'Altura (em centimetros)_y': 'altura_cm'})
df3.head()
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	IMC	altura_cm
0	302	Dados Antropometricos	1.0	2009-03-06	58.00	inf	158.0
1	302	Dados Antropometricos	2.0	2009-01-23	57.00	22.8	158.0
2	302	Dados Antropometricos	3.0	2009-02-06	57.00	22.8	158.0
3	302	Dados Antropometricos	4.0	2009-12-25	62.00	24.8	158.0
4	302	Dados Antropometricos	5.0	2011-07-09	57.75	23.1	158.0

```
[71] # Preenche as linhas em que a mediana do agrupamento do 'Record ID' era NaN
# Utiliza a mediana geral da coluna 'altura_cm' para preencher os dados nulos restantes
mediana = df3['altura_cm'].median()
for i, row in df3.iterrows():
    if pd.isna(row['altura_cm']):
        df3.at[i, 'altura_cm'] = mediana
df3.head()
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	IMC	altura_cm
0	302	Dados Antropometricos	1.0	2009-03-06	58.00	inf	158.0
1	302	Dados Antropometricos	2.0	2009-01-23	57.00	22.8	158.0
2	302	Dados Antropometricos	3.0	2009-02-06	57.00	22.8	158.0
3	302	Dados Antropometricos	4.0	2009-12-25	62.00	24.8	158.0
4	302	Dados Antropometricos	5.0	2011-07-09	57.75	23.1	158.0

Antes e depois de um Record ID que havia outlier:

Antes do tratamento:

```
df3[df3['Record ID'] == 70819]
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	Altura (em centimetros)	IMC
47493	70819	Dados Antropometricos	1.0	2022-02-07	43.50	154.0	18.3
47494	70819	Dados Antropometricos	2.0	2017-07-28	53.50	154.5	22.4
47495	70819	Dados Antropometricos	4.0	2017-08-25	51.20	154.0	21.6
47496	70819	Dados Antropometricos	6.0	2017-09-22	52.30	154.0	22.1
47497	70819	Dados Antropometricos	7.0	2017-10-13	52.30	154.0	22.1
47498	70819	Dados Antropometricos	9.0	2017-10-27	53.00	154.0	22.3
47499	70819	Dados Antropometricos	10.0	2017-11-24	55.00	154.0	23.2
47500	70819	Dados Antropometricos	22.0	2018-02-02	52.90	154.0	22.3
47501	70819	Dados Antropometricos	27.0	2018-06-01	54.50	154.0	23.0
47502	70819	Dados Antropometricos	33.0	2018-11-23	55.35	154.0	23.3
47503	70819	Dados Antropometricos	36.0	2019-05-24	55.00	78.0	90.4
47504	70819	Dados Antropometricos	40.0	2019-11-29	55.50	154.0	23.4
47505	70819	Dados Antropometricos	42.0	2020-01-16	55.35	154.0	23.3
47506	70819	Dados Antropometricos	44.0	2020-02-14	53.80	154.0	22.7
47507	70819	Dados Antropometricos	46.0	2020-03-13	54.55	154.0	23.0
47508	70819	Dados Antropometricos	47.0	2020-07-24	54.55	154.0	23.0
47509	70819	Dados Antropometricos	50.0	2020-11-06	53.40	154.0	22.5

Como é possível visualizar, o padrão de altura desse paciente é 154.0 cm. Porém, no meio desses dados, há um provável erro de digitação onde está preenchido que a altura é 78.0 cm.

Depois do tratamento:

```
df3[df3['Record ID'] == 70819]
```

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	IMC	altura_cm
43926	70819	Dados Antropometricos	1.0	2022-02-07	43.50	18.3	154.0
43927	70819	Dados Antropometricos	2.0	2017-07-28	53.50	22.4	154.0
43928	70819	Dados Antropometricos	4.0	2017-08-25	51.20	21.6	154.0
43929	70819	Dados Antropometricos	6.0	2017-09-22	52.30	22.1	154.0
43930	70819	Dados Antropometricos	7.0	2017-10-13	52.30	22.1	154.0
43931	70819	Dados Antropometricos	9.0	2017-10-27	53.00	22.3	154.0
43932	70819	Dados Antropometricos	10.0	2017-11-24	55.00	23.2	154.0
43933	70819	Dados Antropometricos	22.0	2018-02-02	52.90	22.3	154.0
43934	70819	Dados Antropometricos	27.0	2018-06-01	54.50	23.0	154.0
43935	70819	Dados Antropometricos	33.0	2018-11-23	55.35	23.3	154.0
43936	70819	Dados Antropometricos	36.0	2019-05-24	55.00	90.4	154.0
43937	70819	Dados Antropometricos	40.0	2019-11-29	55.50	23.4	154.0
43938	70819	Dados Antropometricos	42.0	2020-01-16	55.35	23.3	154.0
43939	70819	Dados Antropometricos	44.0	2020-02-14	53.80	22.7	154.0
43940	70819	Dados Antropometricos	46.0	2020-03-13	54.55	23.0	154.0
43941	70819	Dados Antropometricos	47.0	2020-07-24	54.55	23.0	154.0
43942	70819	Dados Antropometricos	50.0	2020-11-06	53.40	22.5	154.0

### 4.2.3. Hipóteses:

Hipótese 1: A morfologia do tumor influencia sobre a escolha do tratamento a ser empregado.

1. Motivação: Estudos

(<https://www.arca.fiocruz.br/handle/icict/4892>)

([http://objdig.ufrj.br/50/teses/m/CCS\\_M\\_MarceloSobralLeite.pdf](http://objdig.ufrj.br/50/teses/m/CCS_M_MarceloSobralLeite.pdf))

Através de pesquisas, foi possível perceber que a morfologia do tumor interfere no tratamento, uma vez que dados mostram que a morfologia tem relação direta com a gravidade do tumor, em critérios como agressividade, proliferação das células cancerígenas, entre outras coisas. Assim, é necessário considerar a morfologia para escolha de tratamento mais adequado.

Hipótese 2: A idade reflete na saúde do paciente para a resposta imunológica positiva com o tratamento.

1. Motivação: estudos

(<https://doutorjairo.uol.com.br/saude-e-longevidade/cancer-de-mama-na-mulher-idosa/>)

(<https://rmmg.org/exportar-pdf/17/v23n1a16.pdf>)

([http://www.ffclrp.usp.br/imagens\\_defesas/31\\_05\\_2010\\_17\\_13\\_48\\_43.pdf](http://www.ffclrp.usp.br/imagens_defesas/31_05_2010_17_13_48_43.pdf))

Conforme as pesquisas cujos links estão acima, a idade pode ter uma influência sobre o tumor, em fatores como agressividade e tamanho, por isso, a hipótese de que a idade por si só pode ter grande influência sobre a escolha de tratamento.

Hipótese 3: Existe uma relação direta entre o IMC e a gravidade do tumor que influencia na escolha do tratamento.

1. Motivação: Estudos (<https://repositorio.uniceub.br/jspui/handle/prefix/13527>)

(<http://repositorio.aee.edu.br/handle/aee/19122>)

Através de pesquisas foi possível perceber que a saúde do paciente está completamente relacionada ao IMC dela, mais precisamente, o IMC afeta na saúde do paciente, tanto IMC muito alto e muito baixo, podem tornar o paciente mais indicado para certo tipo de tratamento por conta dos avanços do tumor e os efeitos colaterais dos tratamentos.

Hipótese 4: A empregabilidade do IMC, com a idade, tem influência sobre a eficácia do modelo.

1. Motivação: Estudos

(<https://repositorio.uniceub.br/jspui/handle/prefix/13527>)

(<http://repositorio.aee.edu.br/handle/aee/19122>)

(<https://doutorjairo.uol.com.br/saude-e-longevidade/cancer-de-mama-na-mulher-idos>

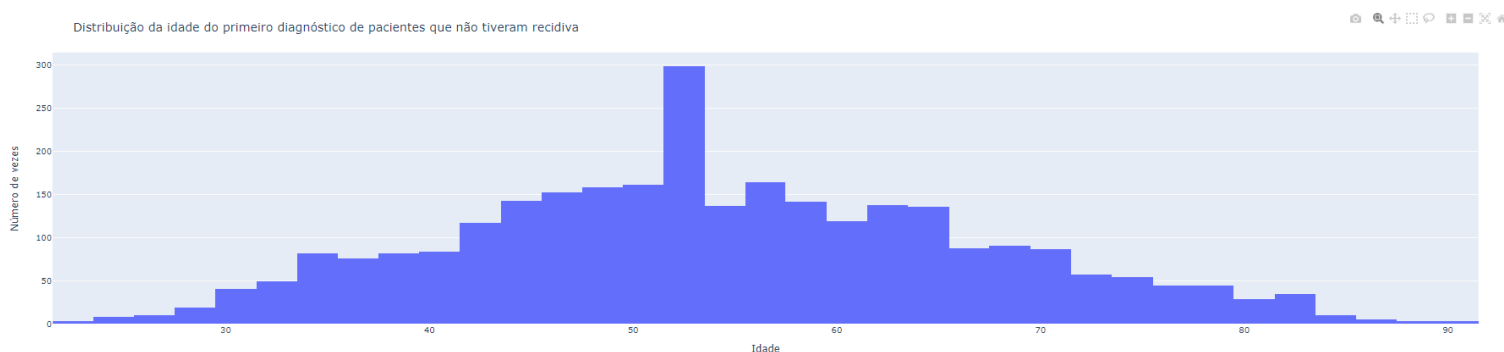
a/)

(<https://rmmg.org/exportar-pdf/17/v23n1a16.pdf>)

([http://www.ffclrp.usp.br/imagens\\_defesas/31\\_05\\_2010\\_17\\_13\\_48\\_43.pdf](http://www.ffclrp.usp.br/imagens_defesas/31_05_2010_17_13_48_43.pdf))

Muitas pesquisas mostram que o IMC e a idade estão correlacionados e, além disso, ambos juntos podem criar mais detalhes cruciais para escolha de tratamento, menos debilitante, ou mais assertivo. Por isso vale considerar a hipótese de empregar os dois juntos nos modelos.

**Hipótese 5: A idade influencia na menor recidiva, maior chance de sucesso, do tratamento.**



1. De acordo com o gráfico, a maioria dos pacientes que se recuperaram descobriram o câncer entre 45 e 65 anos. Dado este que instiga uma dúvida: a base possui mais pessoas de idade avançada ou o índice de recuperação pode estar vinculado com a idade?

As hipóteses foram elaboradas a partir de pesquisas e entrevistas com profissionais da área para criação de hipóteses com uma maior confiabilidade. Além das pesquisas para elaboração de argumentos, foram feitas previsões com as colunas relacionadas com as hipóteses, além da criação de gráficos que demonstram o porquê das hipóteses.

## 4.3. Preparação dos Dados e Modelagem

A modelagem escolhida para o desenvolvimento do projeto, nosso modelo preditivo para auxiliar o médico com o tratamento mais indicado para seu paciente, foi de aprendizagem supervisionada.

### 4.3.1. Modelagem para o problema:

#### 4.3.1.1 Features utilizadas no modelo preditivo:

- Features para realizar os treinos e teste:

grau\_histologico - A literatura médica indica que a histologia do tumor indica a diferenciação das células cancerosas em relação às células normais, dessa forma é possível classificar essas células em diferentes graus que variam de acordo com a agressividade e velocidade de replicação celular das células tumorais. Dessa forma, essa feature ajuda a indicar o tipo de tratamento mais correto, pois dependendo do tamanho e da agressividade, o tratamento pode mudar.

ki67\_maior\_14\_percentage - O ki67 é um tipo de marcador que ajuda a verificar a taxa de replicação celular das células tumorais, dessa forma, a literatura médica estabeleceu que um valor em torno de 14% indica um certo grau elevado dessas replicações, sendo um dos critérios para avaliar a gravidade do tumor e consequentemente influenciando na escolha do tratamento.

subtipo\_tumoral - O subtipo tumoral considera muitos fatores para classificar o tumor, basicamente serve para indicar as características do tumor e assim tendo sua classificação correspondente, podendo ser Luminal A, Luminal B, HER2 positivo e Triplo-Negativo. Os tratamentos podem variar conforme o subtipo tumoral e por isso é uma feature relevante e precisa ser considerada para prever o tratamento correto.

receptor\_de\_progesterona - O receptor hormonal ajuda a indicar o tipo de tratamento para o câncer de mama, no caso ajuda a indicar se é necessário aliar o tratamento com hormonioterapia, se além do receptor de progesterona positivo for constatado receptor de estrogênio positivo.

receptor\_de\_estrogenio - O receptor de estrogênio indica, sozinho ou aliado ao receptor de progesterona, se o tratamento necessitará de hormonioterapia e qual o tratamento mais adequado.

her2\_por\_ihc - O IHC é um método para detecção da proteína HER2, quando essa proteína é detectada, significa que o tumor não é do tipo mais comum e que o tratamento para esse tumor precisa ser feito seguindo o padrão anti-HER2, dessa forma, indica que o tratamento difere do tratamento de câncer HER2 negativo.



her2\_por\_fish - O fish é um exame para detecção da proteína HER2, ele ajuda a indicar se existe a presença dela e no caso, como consequência, ajuda na escolha do tratamento adequado.

grupo\_estadio\_clinico - O grupo de estágio clínico utiliza o padrão da União Internacional Contra o Câncer, ele ajuda a determinar a gravidade do tumor baseando-se em um cálculo a partir de algumas variáveis. Com o estágio clínico determinado, o tratamento é recomendado baseando-se na classificação feita, dessa forma, o grupo de estágio clínico influencia diretamente na escolha do tratamento mais adequado.

classificacao\_tnm\_clinico\_t - A classificação TNM-T indica o tamanho do tumor, impactando diretamente na escolha de terapia adjuvante ou neoadjuvante, pois quando é necessário reduzir o tamanho do tumor, optou-se pela neoadjuvante.

IMC - A literatura mostra que o IMC tem relevância no tratamento do paciente, dessa forma, ao testar no modelo preditivo, pôde-se notar que o acerto aumentou quando o mesmo foi incluído nas Features e por isso foi mantido como fator relevante para escolha do tratamento. O mesmo pode ainda influenciar no sucesso do tratamento.

- Target do modelo preditivo:

regime\_tratamento - O regime de tratamento é a feature que informa qual foi o tratamento escolhido para cada paciente, isso é um dos fatores mais relevantes para o treino e teste do modelo preditivo: ele indica se o resultado apresentado pela máquina está correto ou não e assim, influencia diretamente na determinação de falha ou sucesso do modelo utilizado.

- Features utilizadas para definir as métricas de sucesso do modelo:

com\_recidiva\_regional - No caso de recidiva, é considerado falha no tratamento, uma vez que o tumor não foi tratado corretamente e assim, não houve a eliminação completa, por isso a Feature é de tanta importância, por indicar falha ou sucesso no tratamento.

com\_recidiva\_a\_distancia - No caso de recidiva, é considerado falha no tratamento, uma vez que o tumor não foi tratado corretamente e assim, não houve a eliminação completa, por isso a Feature é de tanta importância, por indicar falha ou sucesso no tratamento.

Independentemente dessa recidiva ser a distância, havendo recidiva, o significado é falha no tratamento.

com\_recidiva\_loca - No caso de recidiva, é considerado falha no tratamento, uma vez que o tumor não foi tratado corretamente e assim, não houve a eliminação completa, por isso a Feature é de tanta importância, por indicar falha ou sucesso no tratamento. A recidiva local indica que o tumor provavelmente não foi extirpado, assim podendo ser considerado tratamento falho.

ultima\_informacao - A última informação do paciente indica o estado dele, por exemplo, se ele está vivo ou morto, isso pode ajudar a definir se o tratamento foi bem-sucedido ou não, pois dependendo do estado do paciente e o porquê, o tratamento pode ter sido falho e a consequência será indicada na feature de última informação.

### 4.3.2. Métricas relacionadas ao modelo:

Foram escolhidas para avaliar os modelos as três métricas seguintes: acurácia do conjunto de treino e de teste, F1-score e matriz de confusão.

A matriz de confusão é importante para a avaliação do modelo porque através dela conseguimos ver o resultado da classificação de cada registro, ela é composta pelos valores: Verdadeiro Positivo, Verdadeiro Negativo, Falso Negativo e Falso Positivo. O Y da matriz são os valores reais e o X da matriz são os valores preditos. No caso da matriz de confusão do nosso projeto, ela é composta por Verdadeiro Tratamento Adjuvante, Verdadeiro Tratamento Neoadjuvante, Falso Tratamento Adjuvante e Falso Tratamento Neoadjuvante, pois nosso modelo não irá prever positivo e negativo.

Figura 1: Demonstração da matriz de confusão

True Positive <b>180</b>	False Negative <b>106</b>
False Positive <b>111</b>	True Negative <b>203</b>

Fonte: [imasters.com.br](http://imasters.com.br)

A análise da matriz de confusão é necessária, também, pois é com os valores demonstrados nela que é possível calcular métricas como acurácia e F1-score.

A acurácia é a divisão da quantidade de acertos no teste, que na matriz de confusão são os valores na diagonal principal, pelo total.

Figura 2: Demonstração do cálculo da acurácia

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

Fonte: [imasters.com.br](http://imasters.com.br)

Já o F1-score é uma média harmônica das métricas Precision e Recall. Para contextualizar, a métrica Precision mostra apenas o valor dos positivos da matriz de confusão, deixando de fora todos os valores considerados negativos, consequentemente deixando de fora alguns positivos por não ter margem de erros, enquanto isso a métrica Recall é exatamente o contrário, ela deixa todos os valores considerados positivos para dentro, o que, em consequência, causa com que alguns valores considerados negativos sejam utilizados. Porém, como o nosso modelo preditivo não é um binário de questão de positivo e negativo, e sim sobre dois tratamentos que podem ser indicados, essas duas métricas, individualmente, não são úteis para ele, por isso escolhemos o F1-score que é uma mistura dos dois e seu objetivo é pegar um pouco dos dois.

Figura 3: Demonstração do cálculo do F1-score

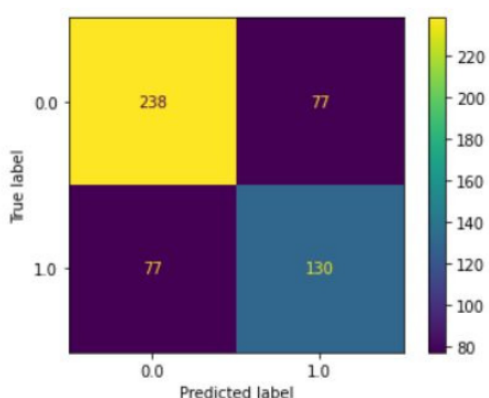
$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fonte: [imasters.com.br](http://imasters.com.br)

Exemplos das métricas em alguns modelos preditivos supervisionados testados:

### Gaussian Naive Bayes

Figura 4: Matriz de confusão do modelo Gaussian Naive Bayes:



Observação: 0.0 se refere a Terapia Adjuvante e 1.0 se refere a Terapia Neoadjuvante

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics

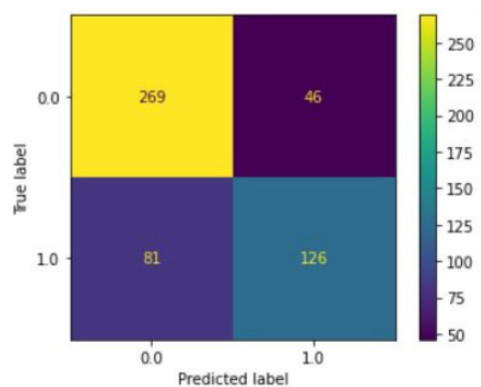
Acurácia Treino: 0.7302631578947368

Acurácia Teste: 0.7049808429118773

F1-score: 0.6280193236714976

### Logistic Regression

Figura 5: Matriz de confusão do modelo Logistic Regression:



Observação: 0.0 se refere a Terapia Adjuvante e 1.0 se refere a Terapia Neoadjuvante

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics

Acurácia Treino: 0.7368421052631579

Acurácia Teste: 0.7567049808429118

F1-score: 0.6649076517150396

### 4.3.3. Primeiro modelo candidato:

O modelo escolhido foi o **Stacking Classifier**, o diferencial dele, para apresentar melhores resultados conforme as métricas utilizadas, é o fato dele usar diferentes modelos como estimadores para obter o melhor resultado, ou seja, ele une o melhor de diferentes modelos, para ter uma maior assertividade. Esses modelos utilizados foram: Random Forest Classifier, SVC, AdaBoostClassifier e Linear SVC.

Figura 6: Código do modelo preditivo Stacking Classifier

```
from sklearn.ensemble import StackingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.ensemble import AdaBoostClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC
from sklearn.metrics import f1_score

estimators = [("rf", RandomForestClassifier(max_depth=4)), ('svr', SVC(max_iter=2000)), ("ada", AdaBoostClassifier())]

n_train = []
n_test = []

alg = StackingClassifier(estimators=estimators, final_estimator=LinearSVC(max_iter=4000))

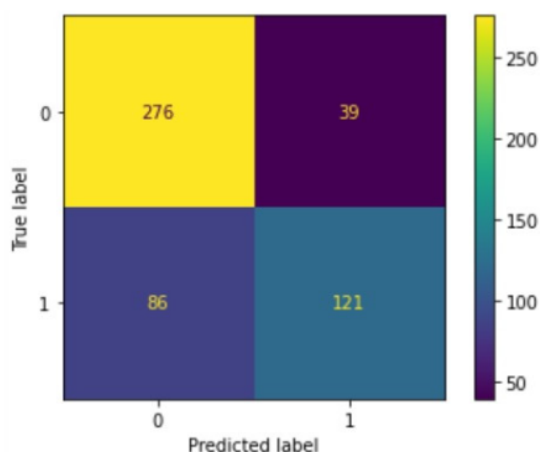
alg.fit(X_train, y_train)
n_train.append(alg.score(X_train, y_train))
n_test.append(alg.score(X_test, y_test))

y_pred1 = alg.predict(X_test)

print(n_train)
print(n_test)
print(f1_score(y_test, y_pred1))
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory

Figura 7: Matriz de confusão do modelo Stacking Classifier



Observação: 0 se refere a Terapia Adjuvante e 1 se refere a Terapia Neoadjuvante

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics

Acurácia Treino: 0.7417763157894737

Acurácia Teste: 0.7605363984674329

F1-score: 0.6594005449591281

Primeiramente, olhando a acurácia, pode-se perceber que não houve **overfitting** e nem **underfitting**, concluindo, então, que ele não ficou muito enviesado nos treinos, o que é considerado um ponto positivo por demonstrar grande margem para melhora da acurácia sem comprometer os treinos. Esse modelo preditivo obteve resultados de testes com maior acurácia em comparação aos outros, foi o único modelo que chegou em 0.76 de acurácia no conjunto de testes, enquanto outros modelos testados ficaram entre 0.70 e 0.75.

Os resultados referentes ao F1-score referem-se ao cálculo feito utilizando tanto precisão quanto o recall para determinar, em um único valor, quão bom está o modelo ao dividir os modelos com margem de erro igualmente para os dois tipos de tratamento. Dessa forma, o resultado do F1-score não está consideravelmente ruim, porém está com grande margem para exploração e melhora, já que é uma métrica importante para o problema onde precisamos compreender o meio termo entre as métricas precision e recall.

Sobre a matriz de confusão do Stacking Classifier, pode-se observar que os valores verdadeiros da diagonal principal da matriz de confusão foram os com maioria de resultados, ou seja, o modelo está acertando o tratamento na maioria das vezes. Porém, pode-se observar que o modelo está acertando na previsão de terapia adjuvante de forma consideravelmente maior do que na terapia Neoadjuvante, onde há maior número de erros. Sendo assim, podemos concluir que ainda é necessário regular melhor o modelo e explorar mais as possibilidades para conseguir reduzir os falsos resultados de neoadjuvantes e adjuvantes que o modelo gera, e entender porque há uma preferência do modelo pelo tipo de tratamento adjuvante, uma vez que estes não são os resultados desejados.

## 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

- Agência Internacional de Pesquisa em Câncer (Iarc): O INCA possui assento tanto no Conselho Diretivo como no Conselho Científico. Gov.br. 2022. Disponível em: <https://www.gov.br/inca/pt-br/aceso-a-informacao/institucional/atuacao-internacional/agencia-internacional-de-pesquisa-em-cancer-iarc> Acesso em: 09 mar. 2023.
- Modelagem preditiva aumenta eficiência de sistemas de saúde: Com ajuda de inteligência artificial, é possível prever riscos e traçar planos de prevenção em sistemas de saúde. Estadão, 2021. Disponível em: <https://summitsaude.estadao.com.br/saude-humanizada/modelagem-preditiva-aumenta-eficiencia-de-sistemas-de-saude/> Acesso em: 09 mar. 2023.
- VIDALE, Giulia. Centro de oncologia do Albert Einstein é eleito o melhor da América Latina: Outros centros brasileiros também integraram a respeitada lista da Newsweek, como o Sírio-Libanês, o A.C. Camargo Câncer Center e a Beneficência Portuguesa. Veja, 2021. Disponível em: Centro de oncologia do Albert Einstein é eleito o melhor da América Latina Leia mais em: <https://veja.abril.com.br/saude/centro-de-oncologia-do-einstein-e-eleito-o-melhor-da-america-latina/> Acesso em: 09 mar. 2023.
- Câncer de Mama Receptor de Hormônio. Oncoguia, 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer-de-mama-receptor-de-hormonio/10879/264/> Acesso em: 09 mar. 2023.
- Câncer de Mama HER2. Oncoguia, 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer-de-mama-her2/10880/264/>. Acesso em: 09 mar. 2023
- ZERWES, Felipe; MILLEN, Eduardo; CAVALCANTE, Francisco Pimentel; NOVITA, Guilherme; FILHO, Hélio Rubens; REIS, João Henrique. Classificações do estadiamento do câncer de mama. câncer de mama Brasil, 2018. Disponível em: <https://www.cancerdemamabrasil.com.br/estadiamento-do-cancer-de-mama/> Acesso em: 09 mar. 2023.
- SATO, Rafael Onuki. Você sabe o que é uma metástase?. Dr. Rafael Onuki Sato, 2018. Disponível em: <https://drrafaelsato.com.br/metastase/#:~:text=Qual%20a%20diferen%C3%A7a%20entre%20recidiva,independente%20do%20local%20de%20aparecimento> Acesso em: 09 mar. 2023.
- TOLENTINO, Grassyara Pinho. Avaliação da composição corporal, qualidade de vida e toxicidade do tratamento quimioterápico em mulheres com câncer de mama. UnB, 2017. Disponível em: <https://repositorio.unb.br/handle/10482/22356> Acesso em: 09 mar. 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.*