

**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 |

**Sumário**

[**1. Introdução**](#_heading=h.2et92p0) **4**

[**2. Objetivos e Justificativa**](#_heading=h.tyjcwt) **5**

[2.1. Objetivos](#_heading=h.3dy6vkm) 5

[2.2. Proposta de Solução](#_heading=h.1t3h5sf) 5

[2.3. Justificativa](#_heading=h.4d34og8) 5

[**3. Metodologia**](#_heading=h.2s8eyo1) **6**

[**4. Desenvolvimento e Resultados**](#_heading=h.17dp8vu) **7**

[4.1. Compreensão do Problema](#_heading=h.3rdcrjn) 7

[4.1.1. Contexto da indústria](#_heading=h.26in1rg) 7

[4.1.2. Análise SWOT](#_heading=h.lnxbz9) 7

[4.1.3. Planejamento Geral da Solução](#_heading=h.35nkun2) 7

[4.1.4. Value Proposition Canvas](#_heading=h.1ksv4uv) 7

[4.1.5. Matriz de Riscos](#_heading=h.44sinio) 7

[4.1.6. Personas](#_heading=h.2jxsxqh) 8

[4.1.7. Jornadas do Usuário](#_heading=h.z337ya) 8

[4.1.8. LGPD](#_heading=h.7s0qxdttaqoz)

[4.2. Compreensão dos Dados](#_heading=h.3j2qqm3) 9

[4.2.1.Exploração dos dados](#_heading=h.2v377qtppdlg)

[4.2.2. Pré processamento](#_heading=h.33bomrb981cw)

[4.2.3. Hipóteses](#_heading=h.v8msiqo8qgxd)

[4.3. Preparação dos Dados e Modelagem](#_heading=h.1y810tw) 10

[4.4. Comparação de Modelos](#_heading=h.4i7ojhp) 11

[4.5. Avaliação](#_heading=h.1ci93xb) 12

[**5. Conclusões e Recomendações**](#_heading=h.3whwml4) **13**

[**6. Referências**](#_heading=h.2bn6wsx) **14**

[**Anexos**](#_heading=h.qsh70q) **15**

# 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 m² 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

*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

### 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.

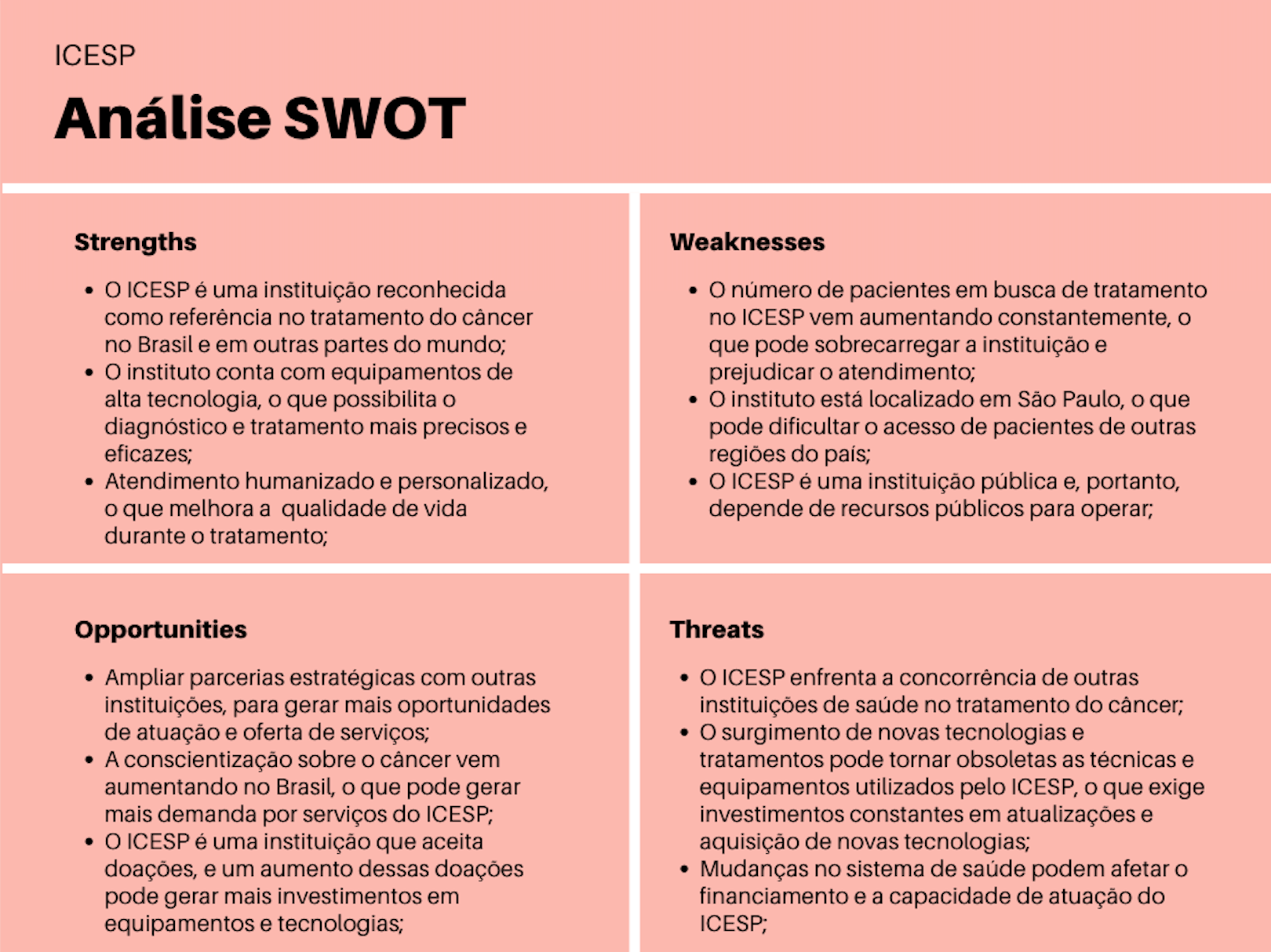
**Ameaça de novos entrantes:** O ICESP possui uma reputação consolidada que pode influenciar para o surgimento de novos entrantes nesse mercado. Porém, a entrada de novos concorrentes pode ser considerada relativamente fácil para instituições de saúde, especialmente se houver uma demanda por serviços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

**

### 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.

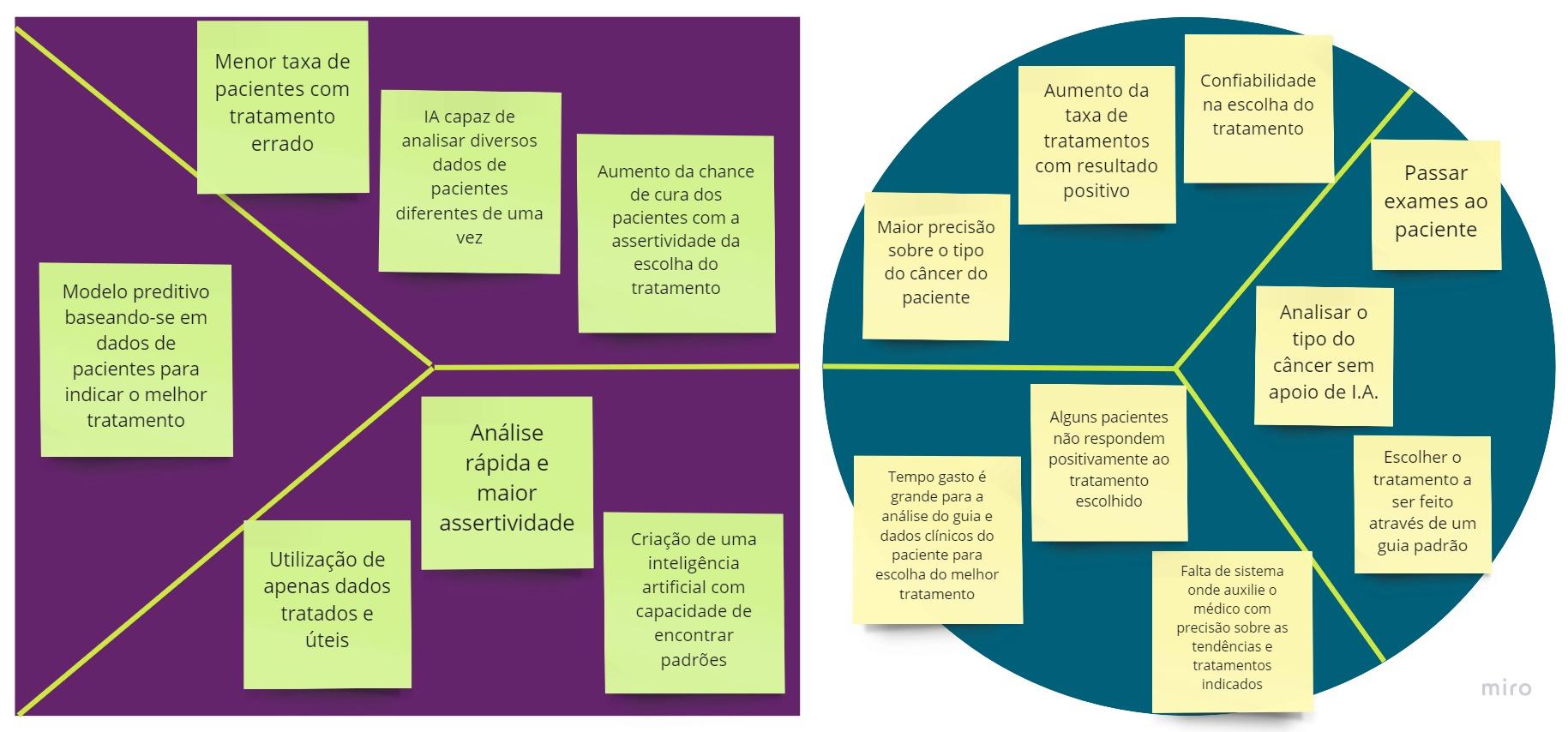
**4.1.3.4. Quais os benefícios trazidos pela solução proposta:**

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.

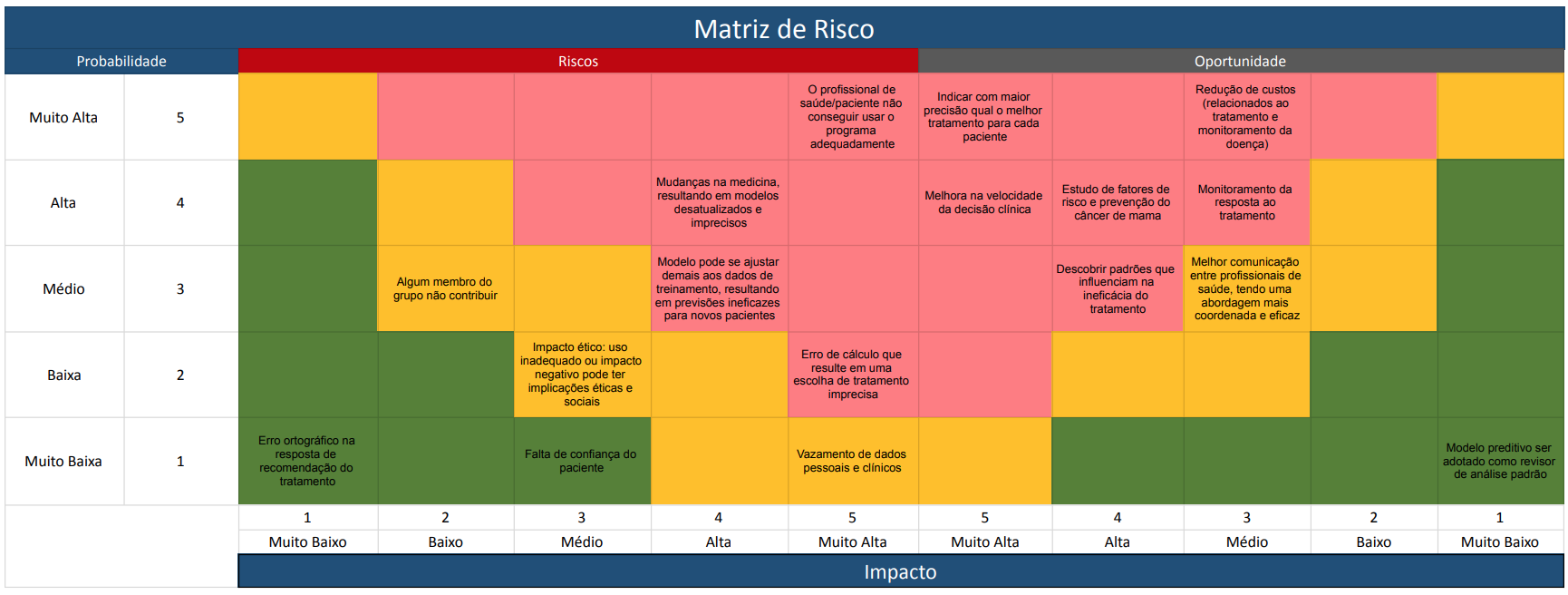
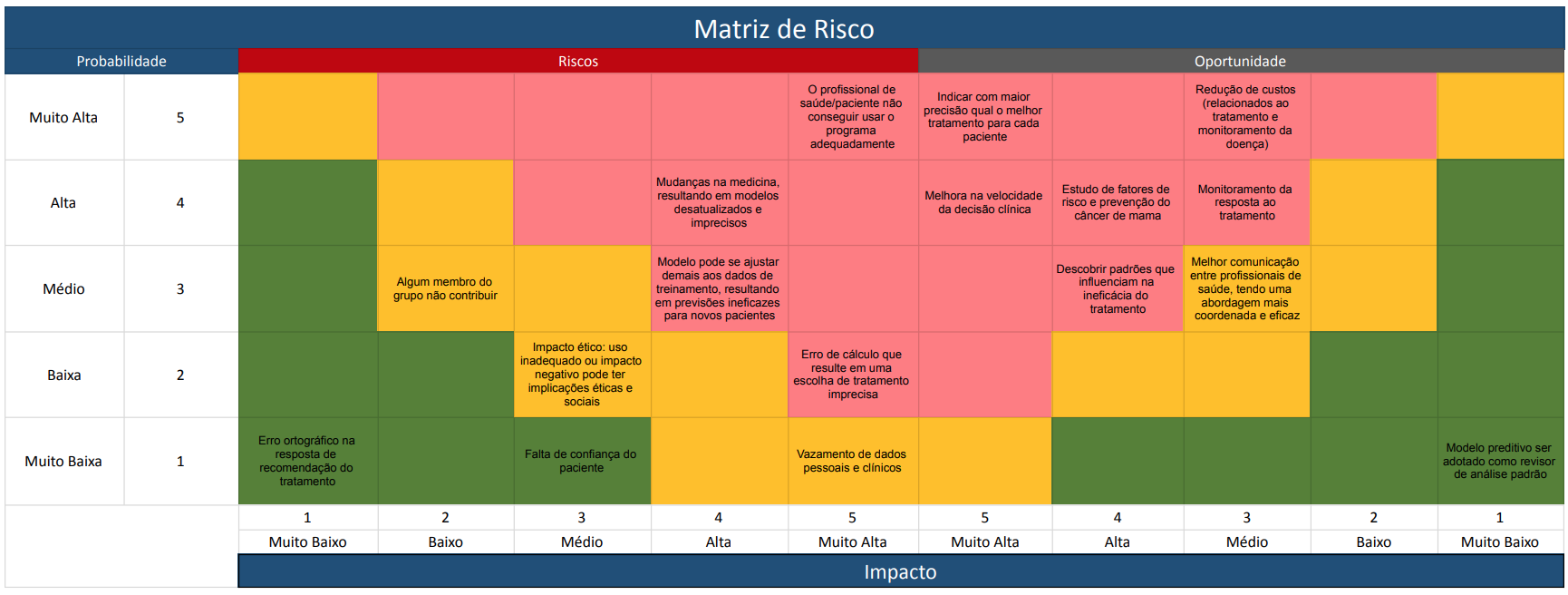
**4.1.3.5. Qual será o critério de sucesso e qual medida será utilizada para o avaliar:**

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

** 

### 

### 

### 

### 

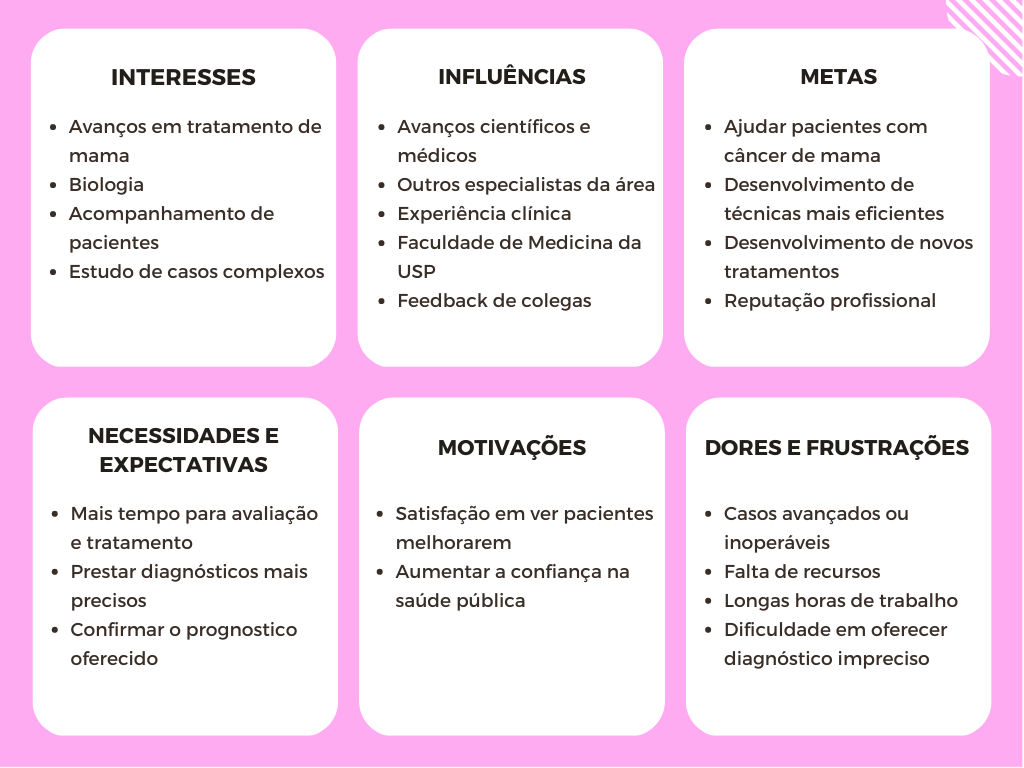
### 

### 

### 

### 4.1.6. Personas

Persona 1 (utiliza o modelo):



Persona 2 (utiliza o modelo):



### 4.1.7. Jornadas do Usuário

*.*

**

### 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:

1. 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 númericas.

Código:

numeric\_cols = merge\_df.select\_dtypes(include=np.number).columns.tolist()

numeric\_cols

Output (nome das colunas númericas):

['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\_O\_1',

'metastase\_ao\_diagnostico\_cid\_O\_2',

'metastase\_ao\_diagnostico\_cid\_O\_3',

'metastase\_ao\_diagnostico\_cid\_O\_4',

'combinacao\_dos\_tratamentos\_realizados\_no\_hospital',

'data\_recidiva',

'local\_recidiva\_a\_xa0\_distancia\_metastase\_1\_cid\_O\_topografia',

'local\_recidiva\_a\_xa0\_distancia\_metastase\_2\_cid\_O\_topografia',

'local\_recidiva\_a\_xa0\_distancia\_metastase\_3\_cid\_O\_topografia',

'local\_recidiva\_a\_xa0\_distancia\_metastase\_4\_cid\_O\_topografia',

'descricao\_da\_morfologia\_de\_acordo\_com\_o\_cid\_O\_cid\_O\_3\_edição)',

'classificacao\_tnm\_patologico\_n',

'classificacao\_tnm\_patologico\_t',

'com\_recidiva\_a\_distancia',

'com\_recidiva\_regional',

'com\_recidiva\_local']

1. 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

1. Gráficos de correlação da estatística descritiva

Gráfico número 1:

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

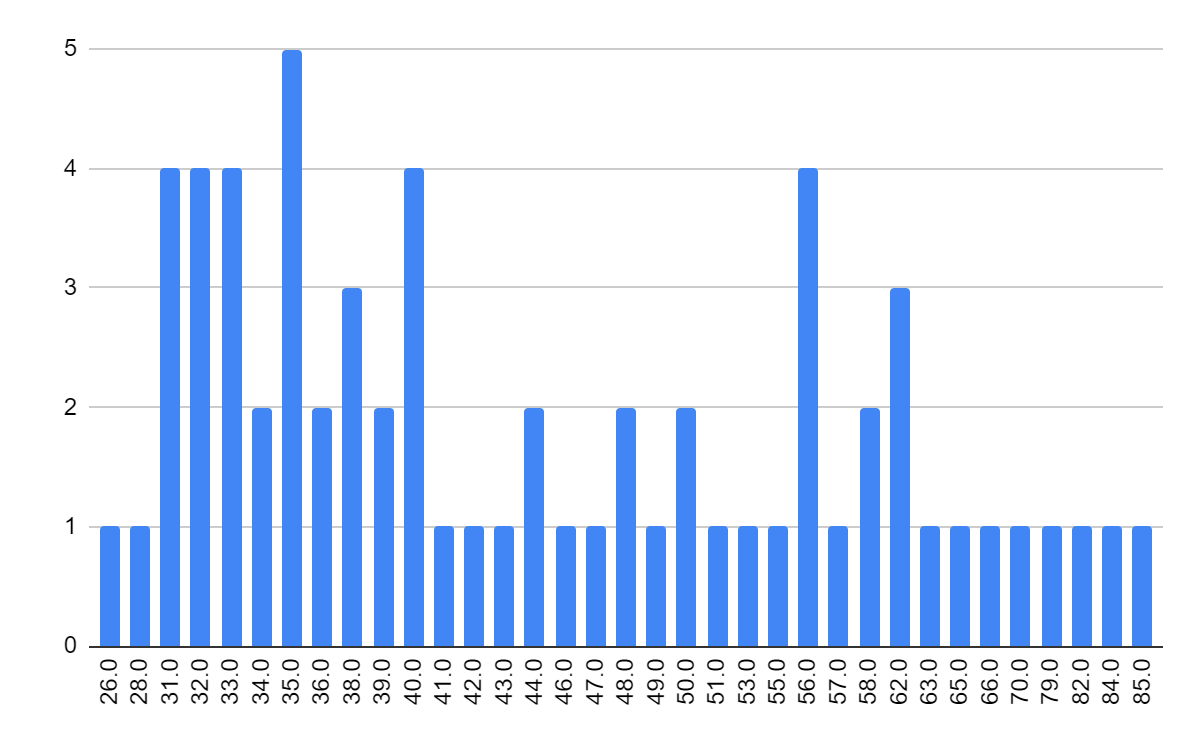
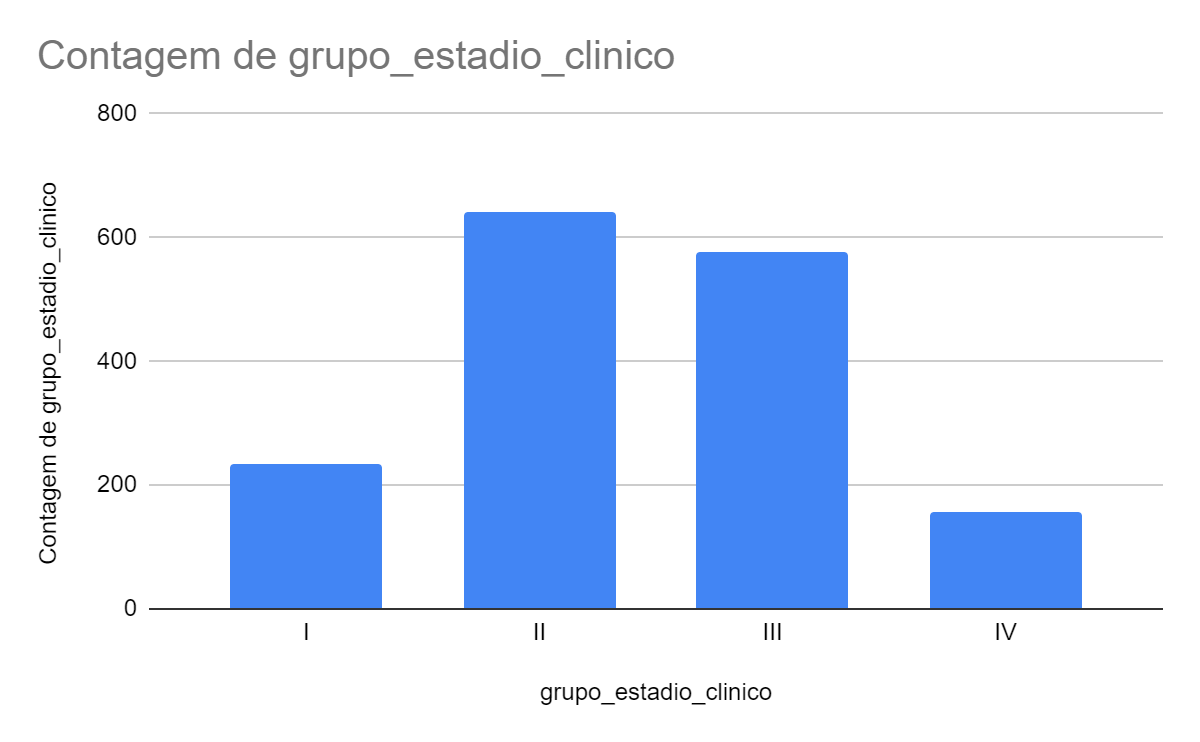


Gráfico que demonstra a relação entre idade e a prática de atividade física (Regular ou frequente) com a idade das pacientes no primeiro diagnóstico, sendo a idade representada no eixo X e a quantidade pacientes que realizam tais atividades no eixo Y.

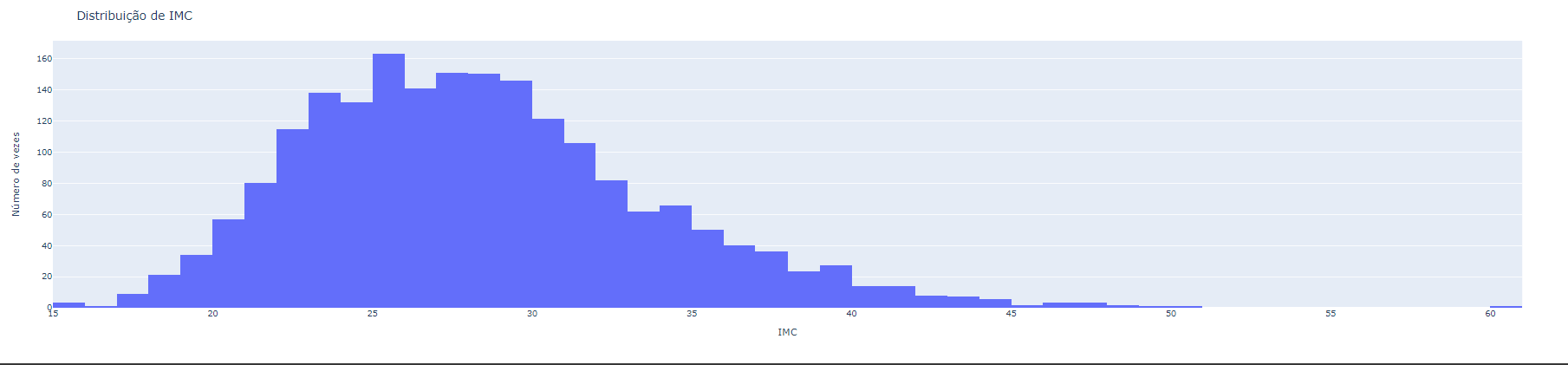
Conclusão: Através deste gráfico podemos notar que pacientes mais jovens têm uma maior tendência a praticar exercícios com uma regularidade ou frequência, talvez por questões hormonais e físicas. Dessa forma vale observar futuramente a relação entre prática de atividades físicas e sobrevivência e cura 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.

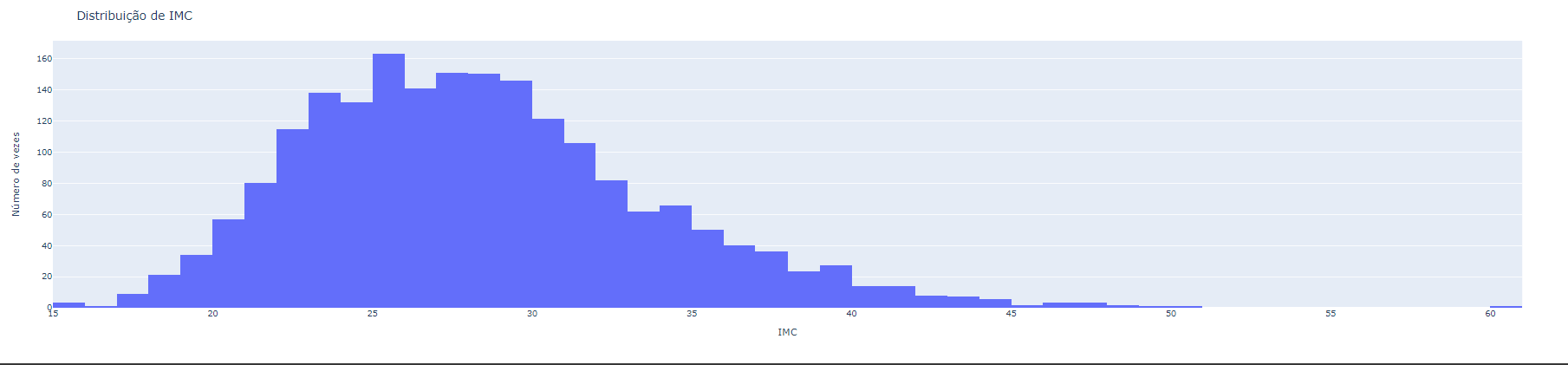
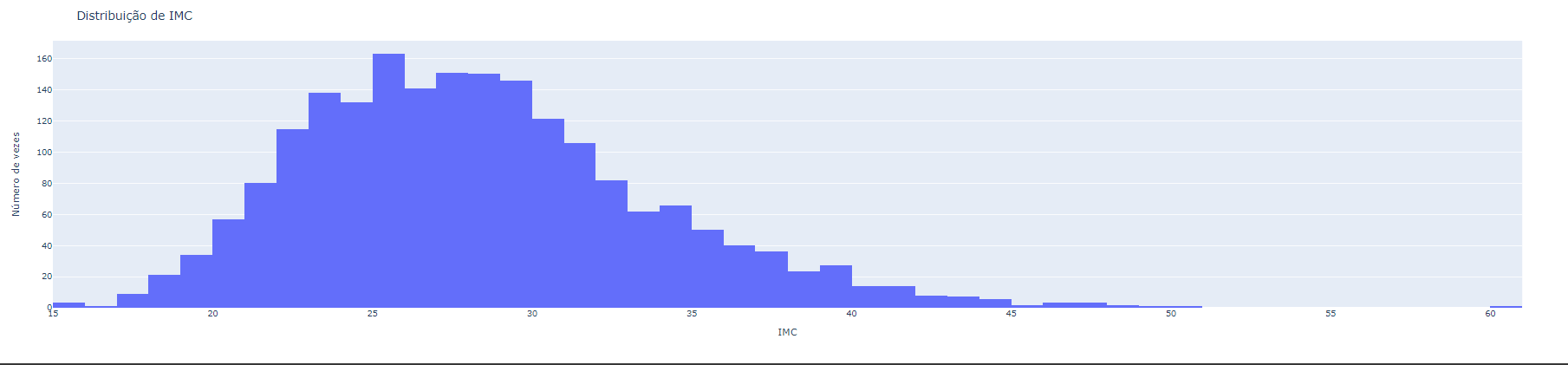
Gráfico número 2:



Conclusão: Segundo o gráfico número 2 percebe-se a quantidade de pacientes em cada grupo de estádio 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:





Conclusão:

Distribuição de IMC: O gráfico demonstra a distribuição do IMC em pacientes com câncer de mama, é possível perceber uma forte tendência de pacientes com sobrepeso e obesidade, representando uma parcela considerável da quantidade de pacientes. A incidência de sobrepeso e obesidade representa a maioria no total de pacientes, dessa forma, é válido observar que os pacientes saudáveis são uma minoria e pode haver uma relação que deve ser testada entre o IMC e a incidência.

Gráfico número 5:

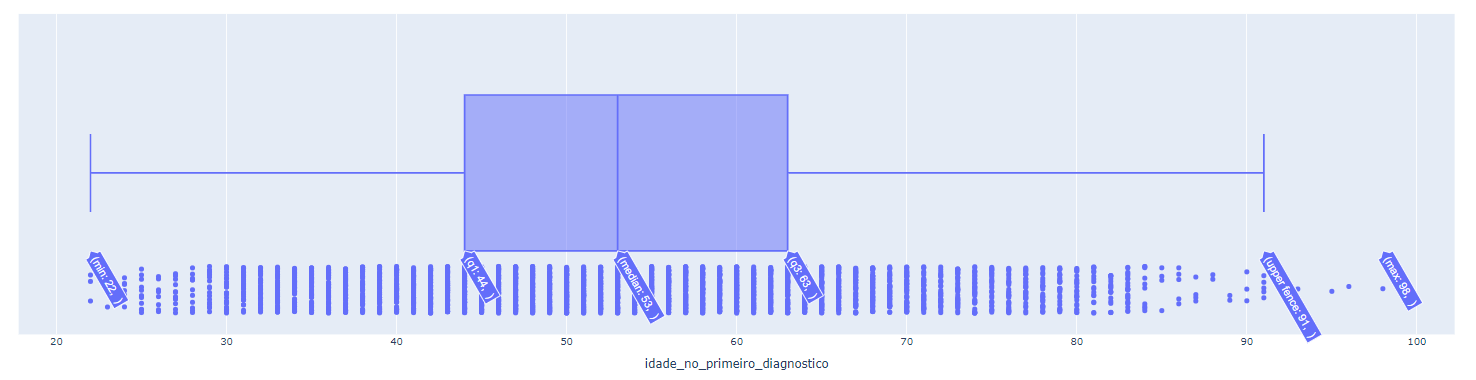


Gráfico número 6:

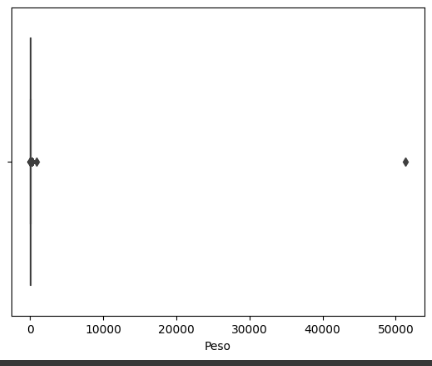
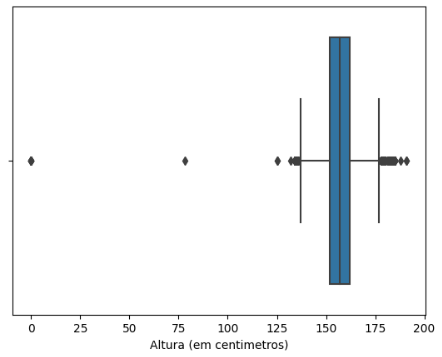


Gráfico número 7:

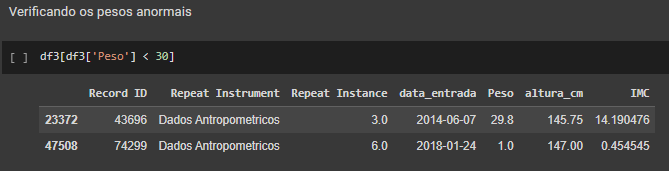


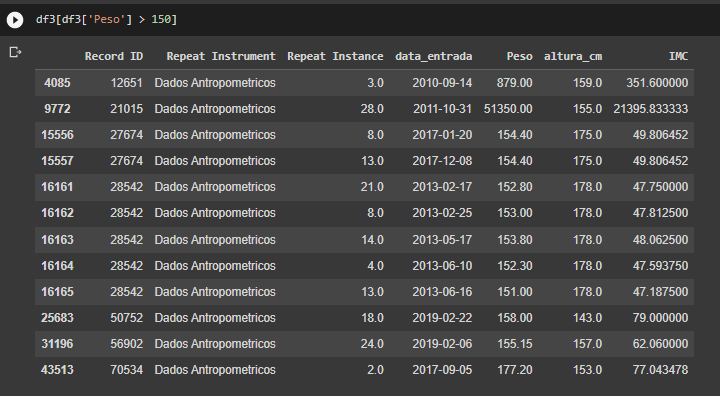
Nos gráficos 5, 6 e 7, estilo boxplot, pode-se identificar a análise descritiva de cada coluna especificada, essas foram: *idade\_no\_primeiro\_diagnostico, Peso* e *Altura (em centimetros).* Podemos observar os principais pontos de uma coluna para começar a limpeza dos dados, esses são: mediana, q1, 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:

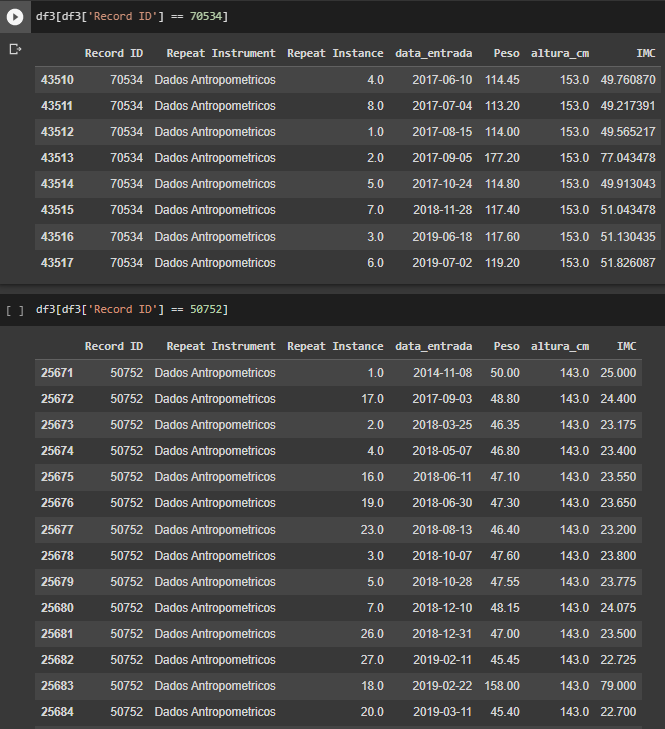
1. Cite quais são os outliers e qual correção será aplicada.

Visualizamos no processo de tratamento dos dados, que havia outliers na coluna *Peso* , *idade\_no\_primeiro\_diagnostico* e *Altura (em centimetros).* 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 imagens com os métodos utilizados:



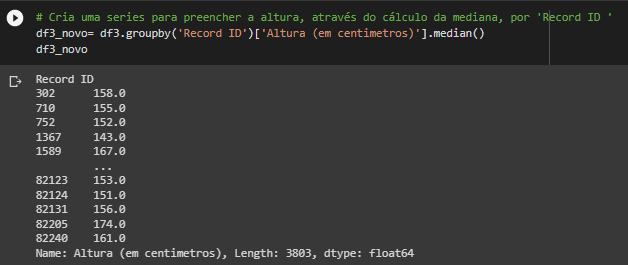


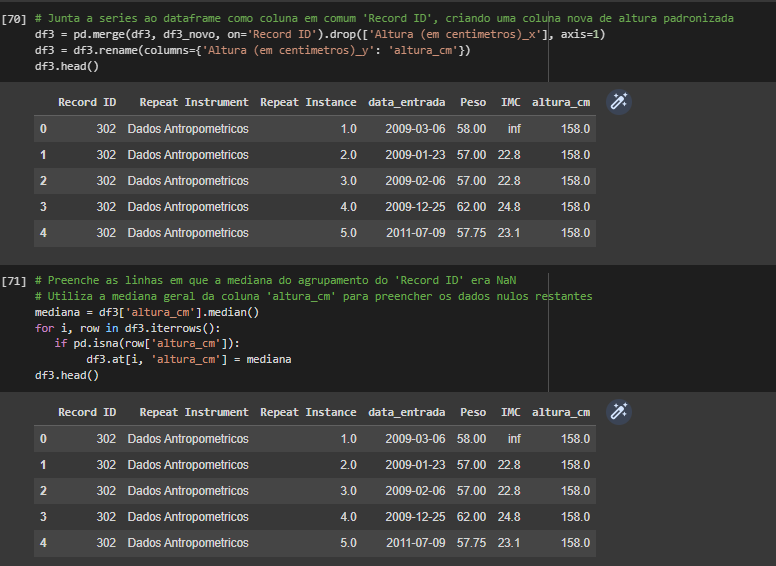
Tivemos como conclusão que os valores 1.00, 879.00 e 51350.00 são claramente outliers providos, provavelmente, de erros de digitação. E então, passamos para a fase de verificar valores levemente acima dos outros, agrupando pelo seu ID e comparando-os , que poderiam, ou não, ser considerados outliers.



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.

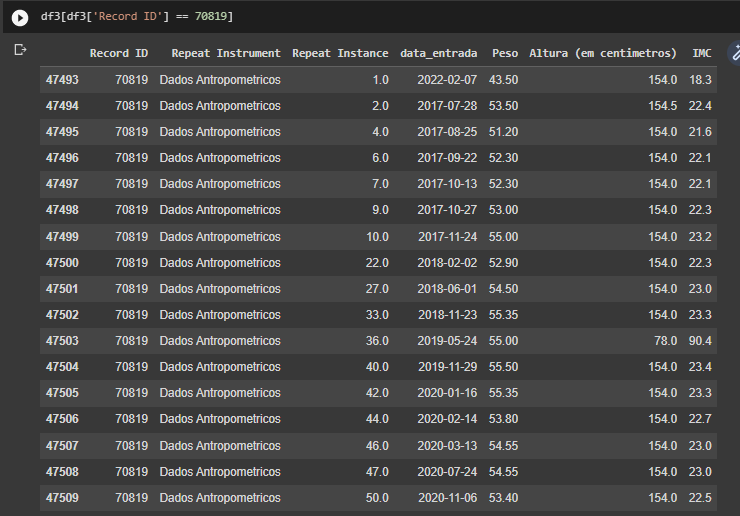
Esse método foi utilizado nas colunas *Peso* e *idade\_no\_primeiro\_diagnostico .* A coluna *Altura (em centimetros),* teve seus outliers tratados juntamente com o preenchimento de missings, onde foi criada uma nova coluna preenchendo novamente as alturas, pegando a mediana do Record ID, assim impedindo que fossem mantidos valores destoantes no mesmo paciente.

**

**

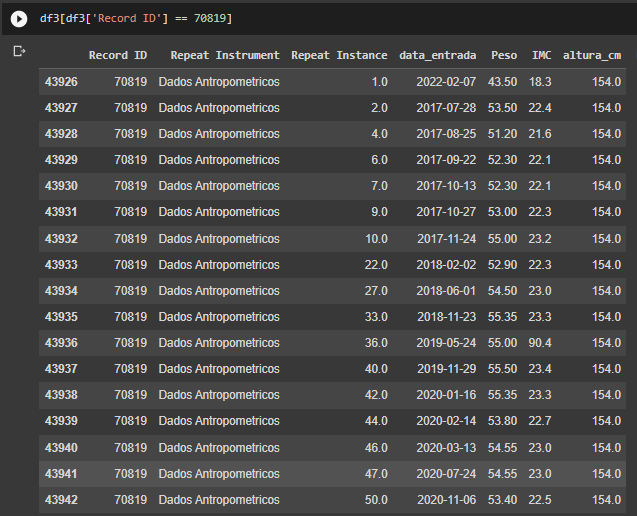
Antes e depois de um Record ID que havia outlier:

Antes do tratamento:



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:



### 4.2.3. Hipóteses:

Hipótese 1:

1. Pergunta motivadora: A morfologia do tumor tem influência sobre a escolha do tratamento a ser empregado?
2. Motivação: Estudos

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

(<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:

1. Pergunta motivadora: a idade reflete na saúde do paciente para a resposta imunológica positiva com o tratamento?
2. 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>)

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:

1. Pergunta motivadora: Existe uma relação direta entre o IMC e a gravidade do tumor que influencia na escolha do tratamento sem aliar com a idade?
2. 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:

1. Pergunta motivadora: A empregabilidade do IMC, com a idade, tem influência sobre a eficácia do modelo?
2. Motivação: Estudos

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

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

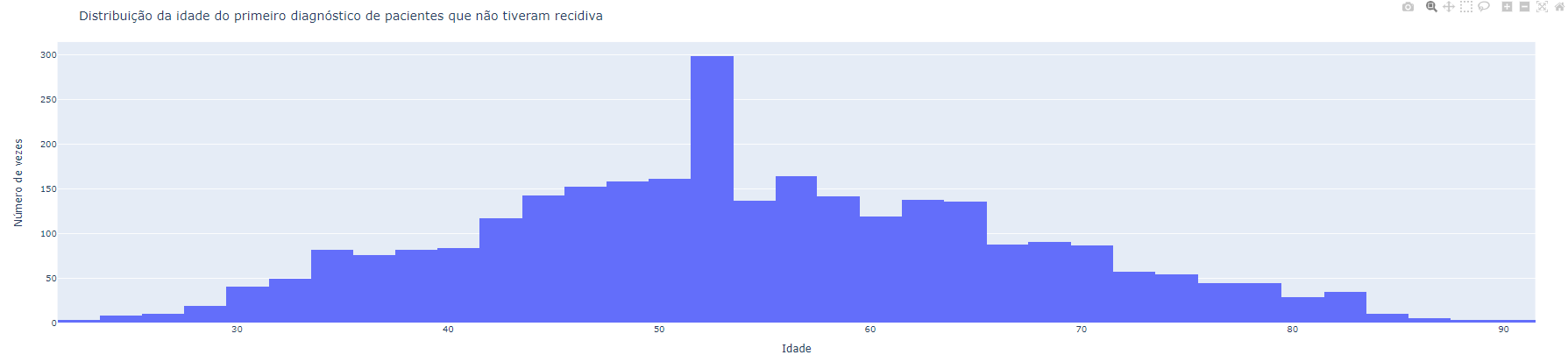
(<https://doutorjairo.uol.com.br/saude-e-longevidade/cancer-de-mama-na-mulher-idosa/>)

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

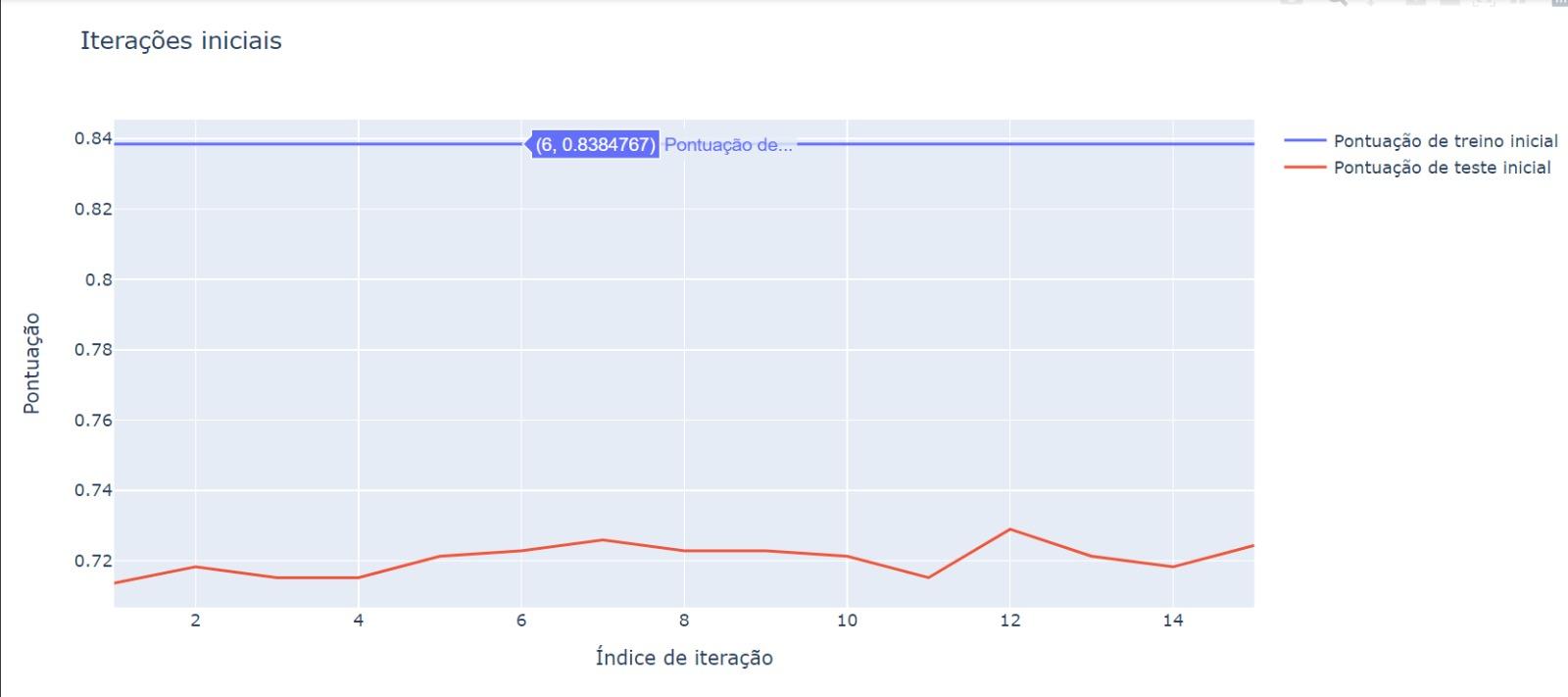
1. Pergunta motivadora: A idade influencia na menor recidiva, maior chance de sucesso, do tratamento?
2. De acordo com o gráfico, a maioria dos pacientes que se recuperaram descobriram o câncer entre 45 e 65 anos. Dado este que instiga uma dúvida: a base possui mais pessoas de idade avançada ou o índice 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 predições com as colunas relacionadas com as hipóteses, além da criação de gráficos que demonstram o porquê das hipóteses.

**Condições de teste inicial:**

1. Colunas utilizadas: ['Ki67\_acima\_14','subtipo\_tumoral',"prog%","estr%","HER2\_IHC",'estadio\_clinico',"TNM-T",'recidiva\_dist','recidiva\_reg','recidiva\_loc',"ultima\_informacao", "regime"]
2. Método utilizado nos testes: Random Forest Classifier (nenhum parâmetro especificado)
3. Separação para o treino: 30% reservado para teste e a seed de separação é 1234
4. Número inicial de dados: 2176
5. Pontuações iniciais:

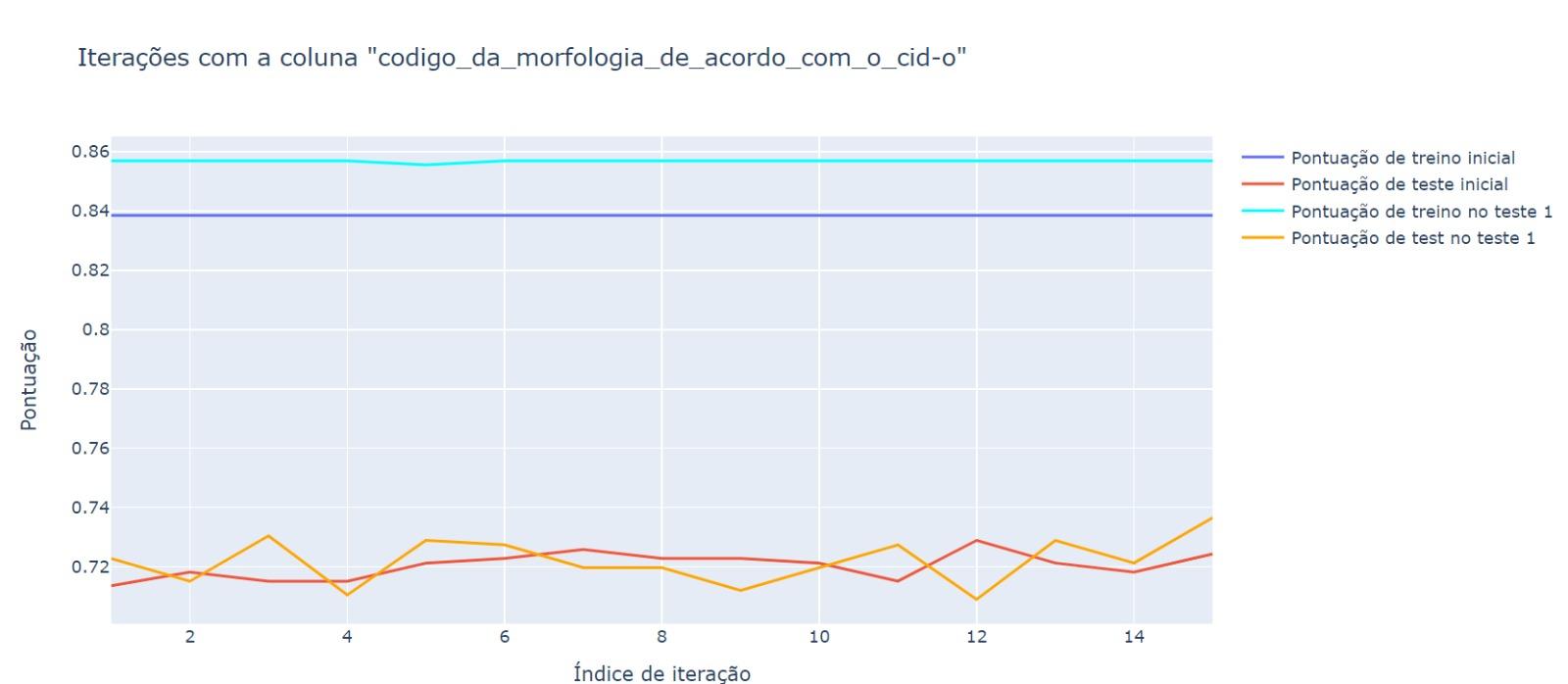
| **nº** | **Pontuação treino** | **Pontuação teste** |
| --- | --- | --- |
| **1** | 0.8384766907419566 | 0.7136294027565084 |
| **2** | 0.8384766907419566 | 0.7182235834609495 |
| **3** | 0.8384766907419566 | 0.7151607963246555 |
| **4** | 0.8384766907419566 | 0.7151607963246555 |
| **5** | 0.8384766907419566 | 0.7212863705972435 |
| **6** | 0.8384766907419566 | 0.7228177641653905 |
| **7** | 0.8384766907419566 | 0.7258805513016845 |
| **8** | 0.8384766907419566 | 0.7228177641653905 |
| **9** | 0.8384766907419566 | 0.7228177641653905 |
| **10** | 0.8384766907419566 | 0.7212863705972435 |
| **11** | 0.8384766907419566 | 0.7151607963246555 |
| **12** | 0.8384766907419566 | 0.7289433384379785 |
| **13** | 0.8384766907419566 | 0.7212863705972435 |
| **14** | 0.8384766907419566 | 0.7182235834609495 |
| **15** | 0.8384766907419566 | 0.7243491577335375 |
| **Med** | **0.8384766907419566** | **0.7205653394589075** |



**Teste da hipótese 1:**

1. Colunas a serem adicionadas: codigo\_da\_morfologia\_de\_acordo\_com\_o\_cid-o
2. Testes:

| **nº** | **Pontuação treino** | **Pontuação teste** |
| --- | --- | --- |
| **1** | 0.8568614576493763 | 0.7228177641653905 |
| **2** | 0.8568614576493763 | 0.7151607963246555 |
| **3** | 0.8568614576493763 | 0.7304747320061256 |
| **4** | 0.8568614576493763 | 0.7105666156202144 |
| **5** | 0.855548260013132 | 0.7289433384379785 |
| **6** | 0.8568614576493763 | 0.7274119448698315 |
| **7** | 0.8568614576493763 | 0.7197549770290965 |
| **8** | 0.8568614576493763 | 0.7197549770290965 |
| **9** | 0.8568614576493763 | 0.7120980091883614 |
| **10** | 0.8568614576493763 | 0.7197549770290965 |
| **11** | 0.8568614576493763 | 0.7274119448698315 |
| **12** | 0.8568614576493763 | 0.7090352220520674 |
| **13** | 0.8568614576493763 | 0.7289433384379785 |
| **14** | 0.8568614576493763 | 0.7212863705972435 |
| **15** | 0.8568614576493763 | 0.7366003062787136 |
| **Med** | **0.8567739111402933** | **0.7220010209290454** |

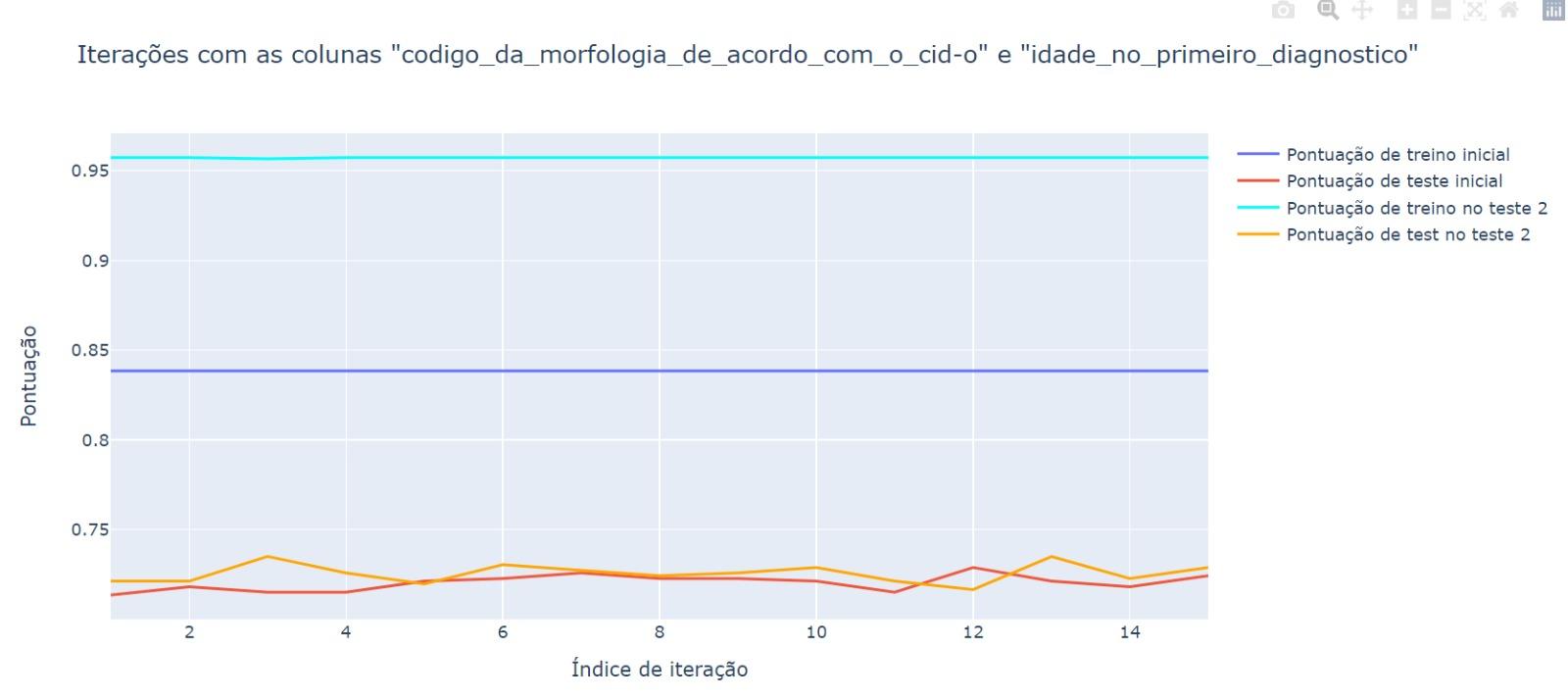


1. Conclusões: Os testes demonstram que existe sim uma relação, mesmo que aparentemente pequena, o código da morfologia acaba influenciando na eficácia do modelo, principalmente nos treinos. Além disso, por ter uma margem pequena de testes, é inconclusivo o impacto que o código da morfologia tem sobre a escolha do tratamento, mas presume-se que ele seja mais positivo do que negativo.
2. Observação: Mantém-se a coluna da morfologia no teste das outras hipóteses para checar se existe uma correlação.

**Teste da hipótese 2:**

1. Colunas a serem adicionadas: idade\_no\_primeiro\_diagnostico e codigo\_da\_morfologia\_de\_acordo\_com\_o\_cid-o
2. Testes:

| **nº** | **Pontuação treino** | **Pontuação teste** |
| --- | --- | --- |
| **1** | 0.9573210768220617 | 0.7212863705972435 |
| **2** | 0.9573210768220617 | 0.7212863705972435 |
| **3** | 0.9566644780039396 | 0.7350689127105666 |
| **4** | 0.9573210768220617 | 0.7258805513016845 |
| **5** | 0.9573210768220617 | 0.7197549770290965 |
| **6** | 0.9573210768220617 | 0.7304747320061256 |
| **7** | 0.9573210768220617 | 0.7274119448698315 |
| **8** | 0.9573210768220617 | 0.7243491577335375 |
| **9** | 0.9573210768220617 | 0.7258805513016845 |
| **10** | 0.9573210768220617 | 0.7289433384379785 |
| **11** | 0.9573210768220617 | 0.7212863705972435 |
| **12** | 0.9573210768220617 | 0.7166921898928025 |
| **13** | 0.9573210768220617 | 0.7350689127105666 |
| **14** | 0.9573210768220617 | 0.7228177641653905 |
| **15** | 0.9573210768220617 | 0.7289433384379785 |
| **Med** | **0.9572773035675204** | **0.7256763654925983** |



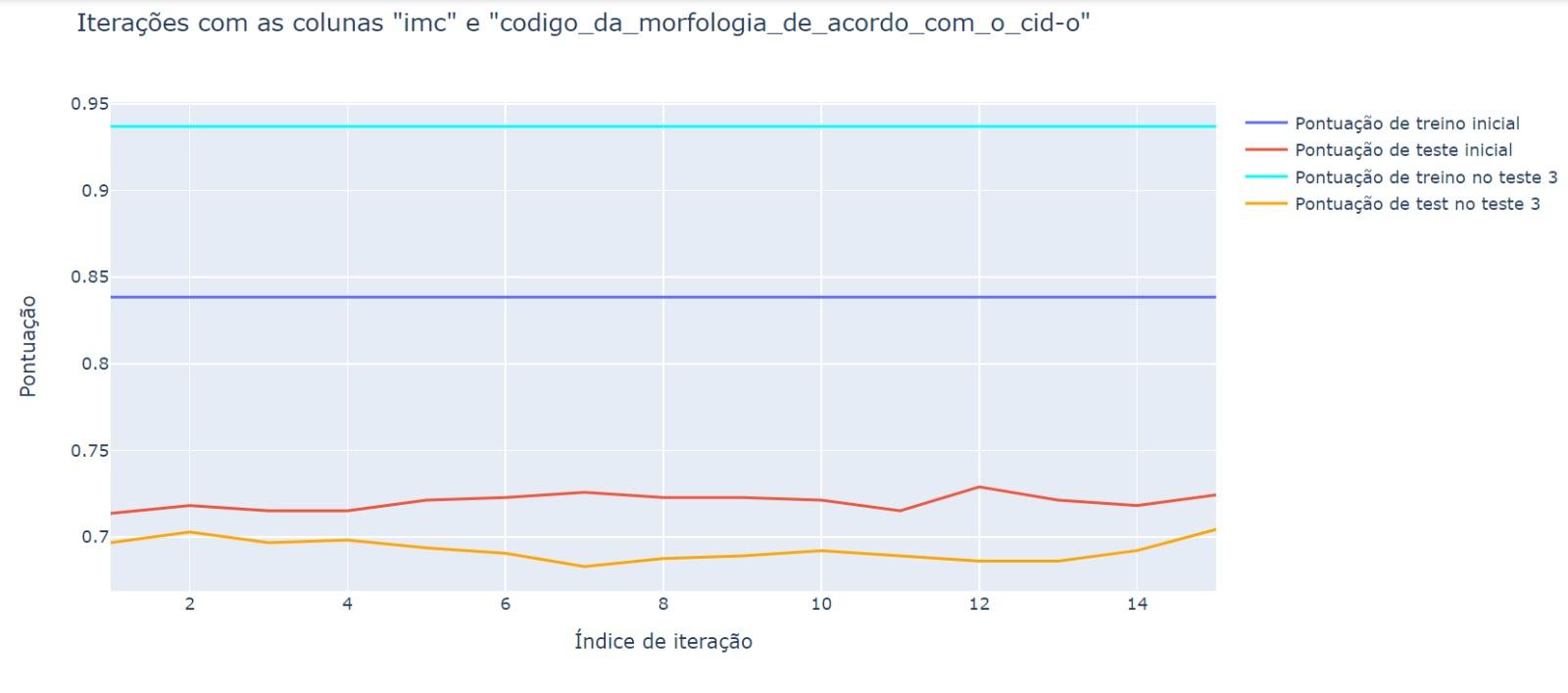
1. Conclusões: o teste demonstrou-se inconclusivo, pois mesmo possuindo uma grande variação na nota do treino, não demonstrou resultados conclusivos na pontuação de teste. Sendo assim, a coluna poderá fornecer informações importantes caso vinculada com alguma outra informação, contudo, por si, não traz mudanças no desempenho.

Pesquisas mostram que idade, aliada com IMC, podem ser fatores determinantes para a gravidade do câncer, por conta disso, o teste foi rodado previamente apenas com idade, para ver se algum fator em particular demonstraria um grande impacto.

**Teste da hipótese 3:**

1. Colunas a serem adicionadas: imc e codigo\_da\_morfologia\_de\_acordo\_com\_o\_cid-o
2. Testes:

| **nº** | **Pontuação treino** | **Pontuação teste** |
| --- | --- | --- |
| **1** | 0.9369665134602758 | 0.6967840735068913 |
| **2** | 0.9369665134602758 | 0.7029096477794793 |
| **3** | 0.9369665134602758 | 0.6967840735068913 |
| **4** | 0.9369665134602758 | 0.6983154670750383 |
| **5** | 0.9369665134602758 | 0.6937212863705973 |
| **6** | 0.9369665134602758 | 0.6906584992343032 |
| **7** | 0.9369665134602758 | 0.6830015313935681 |
| **8** | 0.9369665134602758 | 0.6875957120980092 |
| **9** | 0.9369665134602758 | 0.6891271056661562 |
| **10** | 0.9369665134602758 | 0.6921898928024502 |
| **11** | 0.9369665134602758 | 0.6891271056661562 |
| **12** | 0.9369665134602758 | 0.6860643185298622 |
| **13** | 0.9369665134602758 | 0.6860643185298622 |
| **14** | 0.9369665134602758 | 0.6921898928024502 |
| **15** | 0.9369665134602758 | 0.7044410413476263 |
| **Med** | **0.9369665134602758** | **0.6925982644206228** |

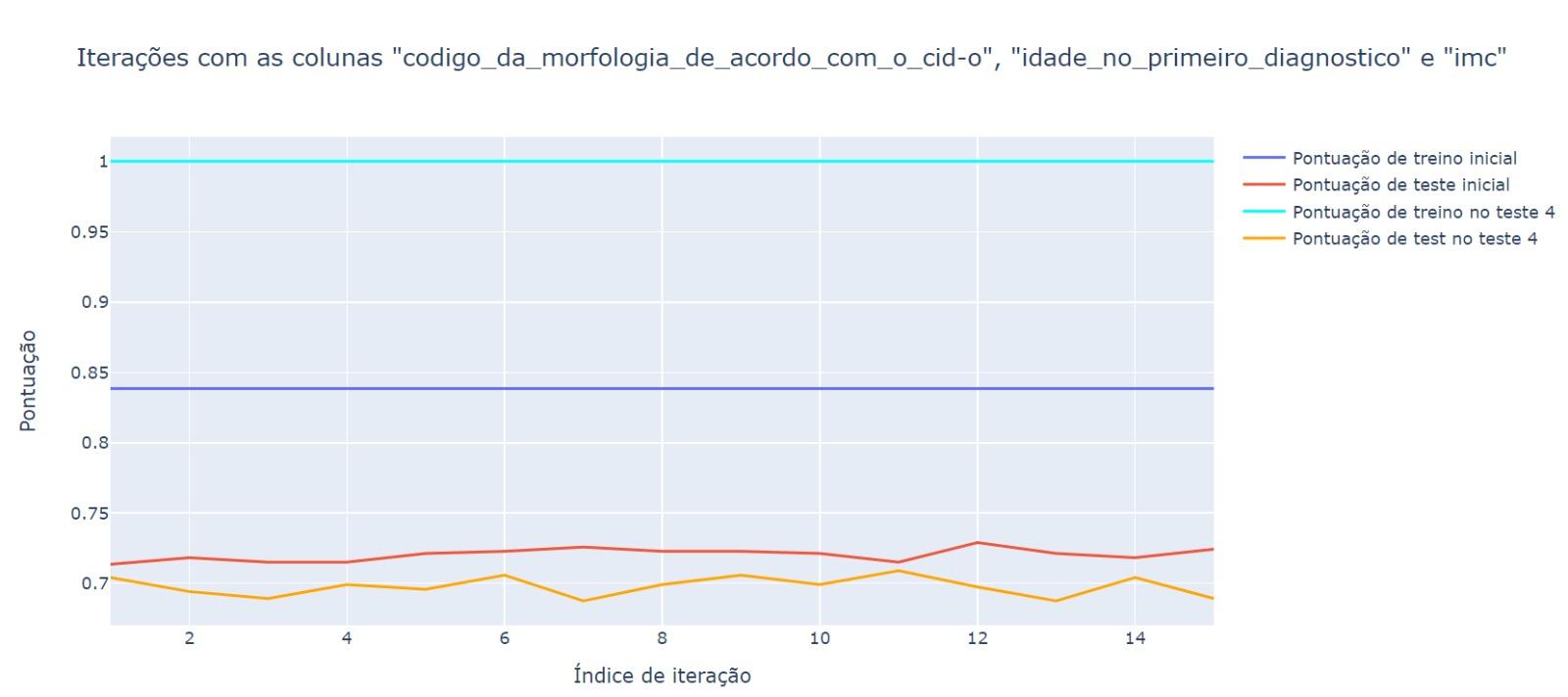


1. Conclusões: Apesar de aparecer um leve declínio na taxa de acertos quando é colocada a variável IMC, as resoluções ainda são inconclusivas, uma vez que estudos demonstram a relação entre tumor e IMC. Existe a necessidade de avaliação futura se o IMC aliado com o fator idade pode ser um indicativo prudente que aumente a precisão do modelo. Para a base de treinamento, o IMC parece ter aumentado a eficácia, mas resta testar mais para compreender melhor seus efeitos quando utilizados para testes.

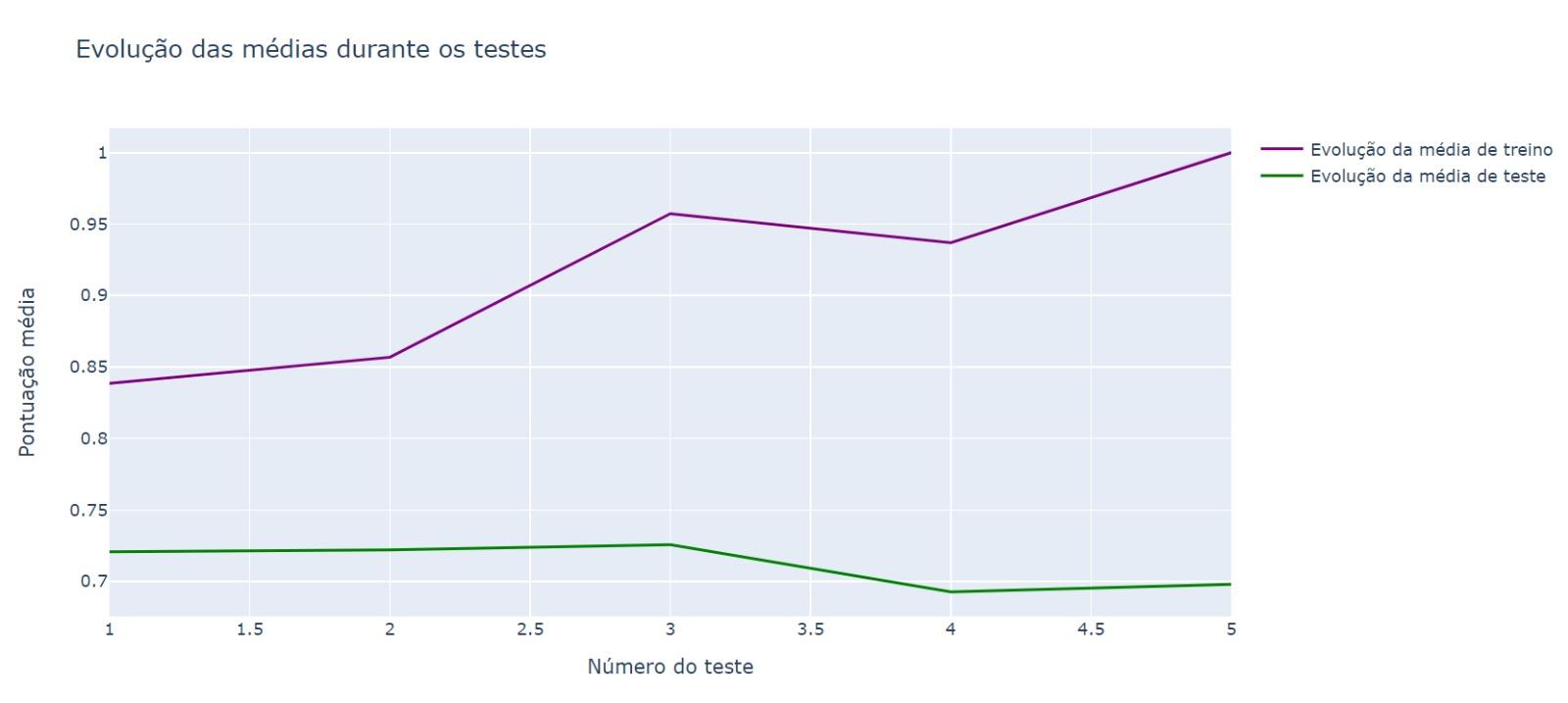
**Teste da hipótese 4:**

1. Colunas a serem adicionadas: imc, idade\_no\_primeiro\_diagnostico e codigo\_da\_morfologia\_de\_acordo\_com\_o\_cid-o
2. Testes:

| **nº** | **Pontuação treino** | **Pontuação teste** |
| --- | --- | --- |
| **1** | 1.0 | 0.7041322314049587 |
| **2** | 1.0 | 0.6942148760330579 |
| **3** | 1.0 | 0.6892561983471074 |
| **4** | 1.0 | 0.6991735537190082 |
| **5** | 1.0 | 0.6958677685950413 |
| **6** | 1.0 | 0.7057851239669422 |
| **7** | 1.0 | 0.687603305785124 |
| **8** | 1.0 | 0.6991735537190082 |
| **9** | 1.0 | 0.7057851239669422 |
| **10** | 1.0 | 0.6991735537190082 |
| **11** | 1.0 | 0.7090909090909091 |
| **12** | 1.0 | 0.6975206611570248 |
| **13** | 1.0 | 0.687603305785124 |
| **14** | 1.0 | 0.7041322314049587 |
| **15** | 1.0 | 0.6892561983471074 |
| **Med** | **1.0** | **0.6978512396694215** |



1. Conclusões: O teste mostrou que quando combinado a idade e o IMC com as colunas iniciais, a certeza dos treinos utilizando a IA random forest chega a 100%, demonstrando algo surpreendente, percebido com estudos feitos pelos membros do grupo. Vale continuar observando mais os dados, uma vez que a certeza de treinos pode variar dependendo da IA escolhida, o treino foi muito enriquecedor, embora os testes tenham sido inconclusivos, por conta das diferenças serem muito pequenas com a amostragem inicial. A sugestão para a melhora do desempenho é o teste de outro framework de IA.



## 

## 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 predições.*

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

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

# 5. Conclusões e Recomendações

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

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

# 6. Referências

[**https://www.gov.br/inca/pt-br/acesso-a-informacao/institucional/atuacao-internacional/agencia-internacional-de-pesquisa-em-cancer-iarc**](https://www.gov.br/inca/pt-br/acesso-a-informacao/institucional/atuacao-internacional/agencia-internacional-de-pesquisa-em-cancer-iarc)

[**https://summitsaude.estadao.com.br/saude-humanizada/modelagem-preditiva-aumenta-eficiencia-de-sistemas-de-saude/**](https://summitsaude.estadao.com.br/saude-humanizada/modelagem-preditiva-aumenta-eficiencia-de-sistemas-de-saude/)

[**https://veja.abril.com.br/saude/centro-de-oncologia-do-einstein-e-eleito-o-melhor-da-america-latina/#:~:text=Em%202021%2C%20o%20Centro%20de,mundial%20de%20melhores%20hospitais%20oncol%C3%B3gicos**](https://veja.abril.com.br/saude/centro-de-oncologia-do-einstein-e-eleito-o-melhor-da-america-latina/#:~:text=Em%202021%2C%20o%20Centro%20de,mundial%20de%20melhores%20hospitais%20oncol%C3%B3gicos)

[**https://www.pravaler.com.br/melhores-faculdades-de-medicina/**](https://www.pravaler.com.br/melhores-faculdades-de-medicina/)

[**https://ensinoepesquisa.icesp.org.br/pt/?\_ga=2.99327704.791521092.1675275308-211678864.1675275308&\_gl=1\*cuwsb5\*\_ga\*MjExNjc4ODY0LjE2NzUyNzUzMDg.\*\_ga\_M9PTNL86PX\*MTY3NTI3NTMwNy4xLjAuMTY3NTI3NTMwNy4wLjAuMA**](https://ensinoepesquisa.icesp.org.br/pt/?_ga=2.99327704.791521092.1675275308-211678864.1675275308&_gl=1*cuwsb5*_ga*MjExNjc4ODY0LjE2NzUyNzUzMDg.*_ga_M9PTNL86PX*MTY3NTI3NTMwNy4xLjAuMTY3NTI3NTMwNy4wLjAuMA)

[**https://fei.edu.br/sites/sicfei/2015/Produ%C3%A7%C3%A3o/O%20SETOR%20DE%20EQUIPAMENTOS%20M%C3%89DICO-HOSPITALARES%20BRASILEIRO%20-%20INOVA%C3%87%C3%83O,%20COMPETITIVIDADE%20E%20DESAFIOS.pdf**](https://fei.edu.br/sites/sicfei/2015/Produ%C3%A7%C3%A3o/O%20SETOR%20DE%20EQUIPAMENTOS%20M%C3%89DICO-HOSPITALARES%20BRASILEIRO%20-%20INOVA%C3%87%C3%83O,%20COMPETITIVIDADE%20E%20DESAFIOS.pdf)

# Anexos

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