OncoAI
Faculdade de Medicina da USP

inteli
instituto
de tecnologia
e liderança

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

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

ICESP

Análise SWOT

Strengths

- O ICESP 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 ICESP 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 ICESP 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 ICESP;
- O ICESP é uma instituição que aceita doações, e um aumento dessas doações pode gerar mais investimentos em equipamentos e tecnologias;

Threats

- O ICESP 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 ICESP, 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 ICESP;

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.

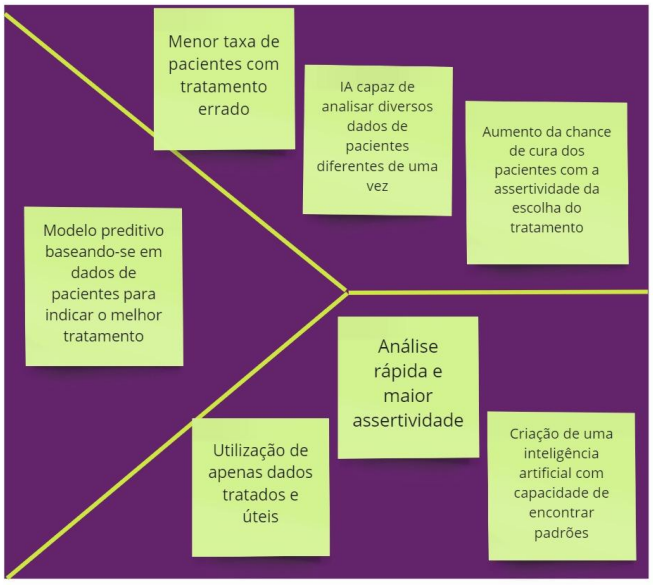
4.1.3.4. Quais os benefícios trazidos pela solução proposta:

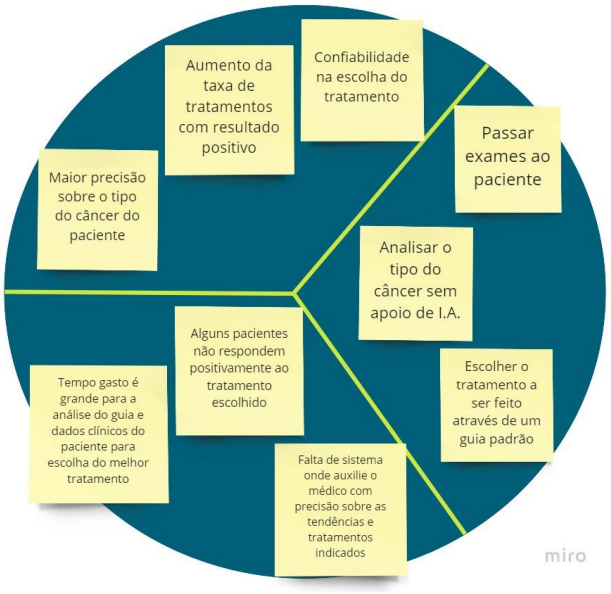
Os principais benefícios são o aumento da posição do médico no diagnóstico e escolha do tratamento, pelo auxílio dos resultados vindos da IA. Através desse auxílio, o médico terá maior chance de ter indicado o melhor tratamento para aquele caso. Sendo assim, aumentará a taxa de pacientes com possível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

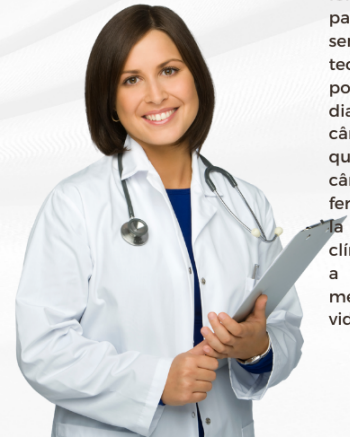
Oportunidade				
Indicar com maior precisão qual o melhor tratamento para cada paciente		Redução de custos (relacionados ao tratamento e monitoramento da doença)		
Melhora na velocidade da decisão clínica	Estudo de fatores de risco e prevenção do câncer de mama	Monitoramento da resposta ao tratamento		
	Descobrir padrões que influenciam na ineficácia 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 prognostico 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
- Sedentaria
- Insegura
- Corajosa

INTERESSES

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

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

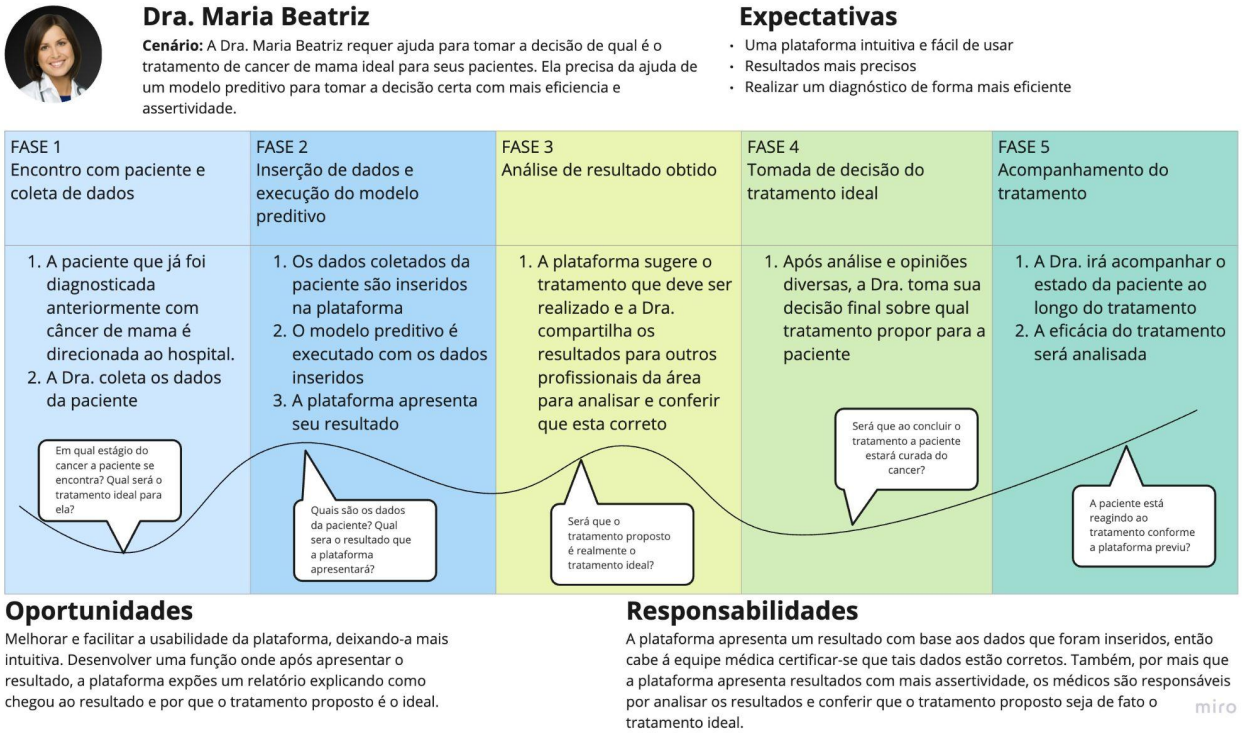
MOTIVAÇÕES

- 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



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 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']

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

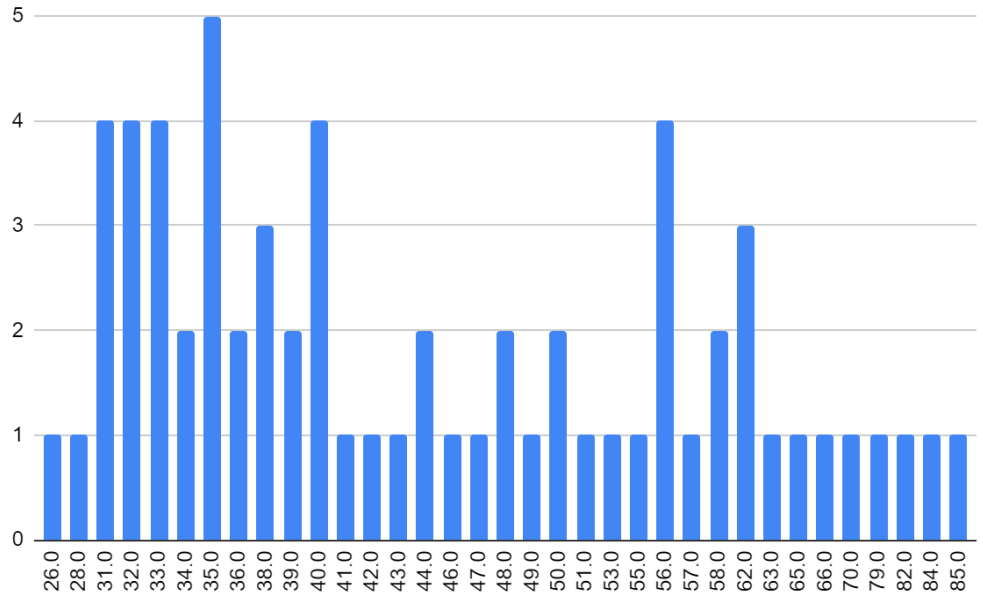
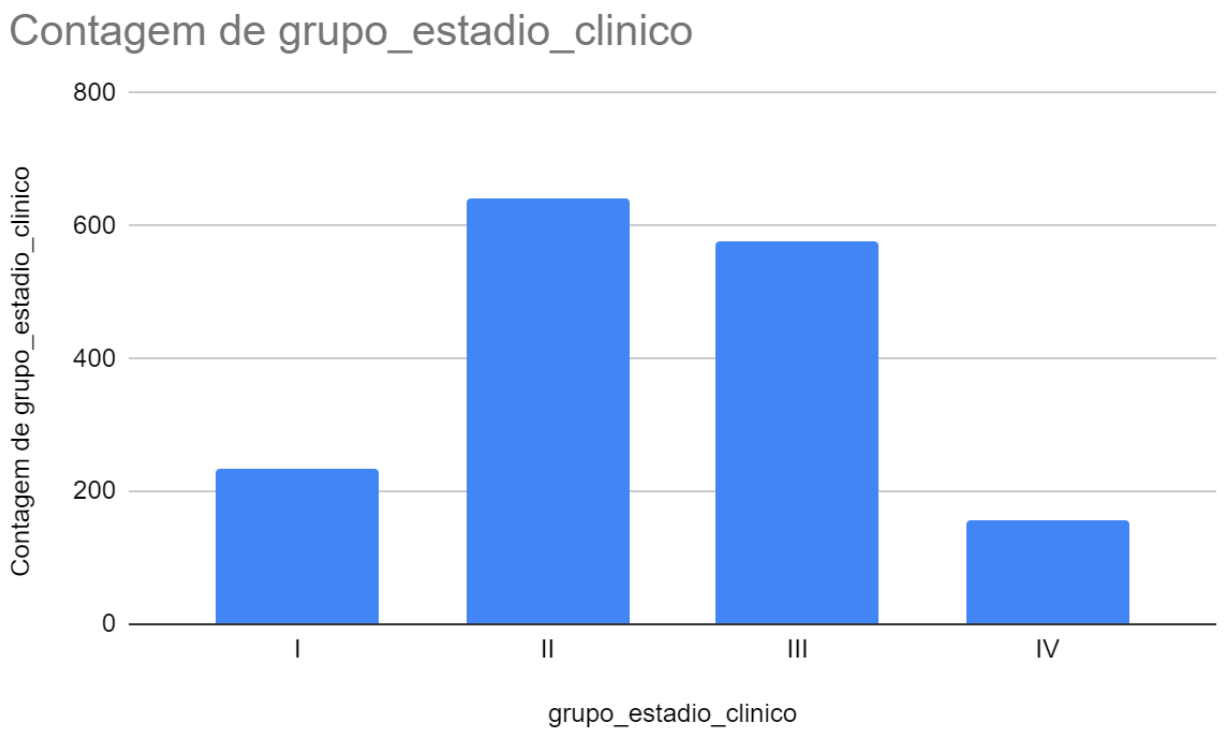


Gráfico que demonstra a relação entre idade e a prática de atividade física (Regular ou frequente) com a idade das pacientes no primeiro diagnóstico, sendo a idade representada no eixo X e a quantidade pacientes que realizam tais atividades no eixo Y.

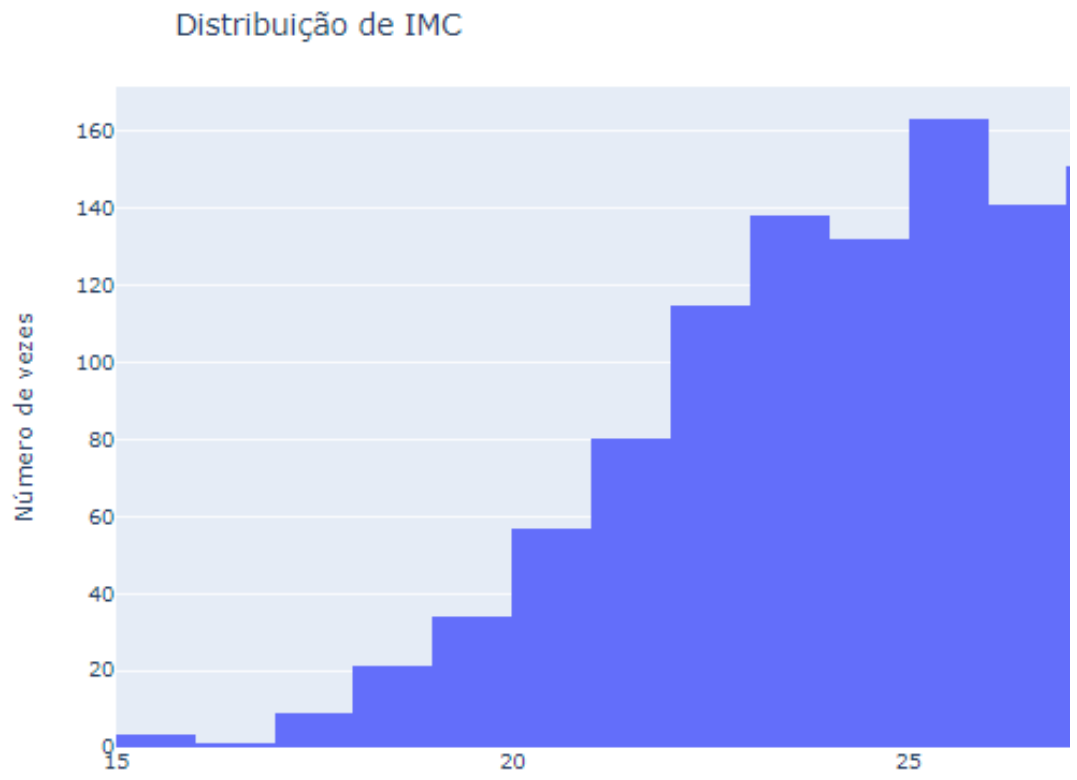
Conclusão: Através deste gráfico podemos notar que pacientes mais jovens têm uma maior tendência a praticar exercícios com uma regularidade ou frequência, talvez por questões hormonais e físicas. Dessa forma vale observar futuramente a relação entre prática de atividades físicas e sobrevivência e cura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

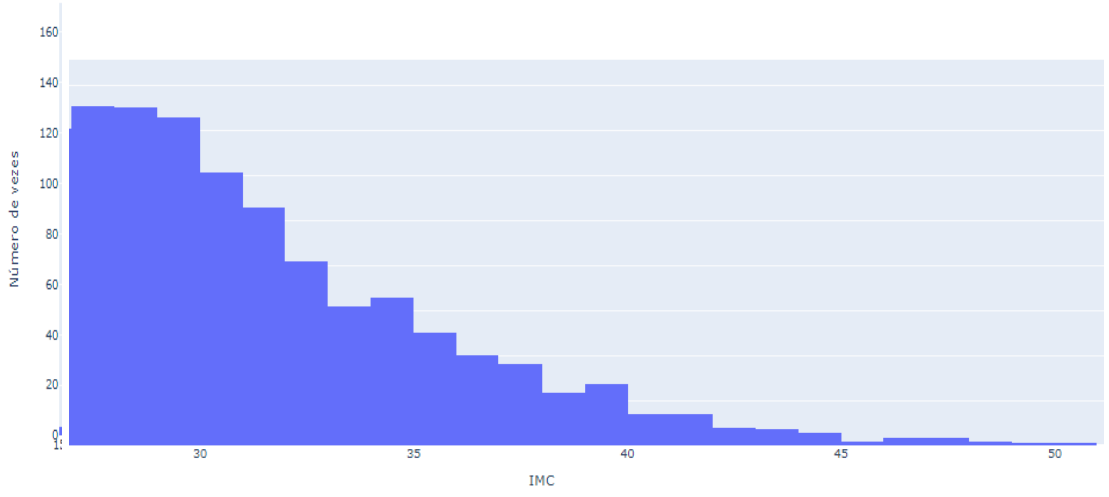
Gráfico número 2:



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:

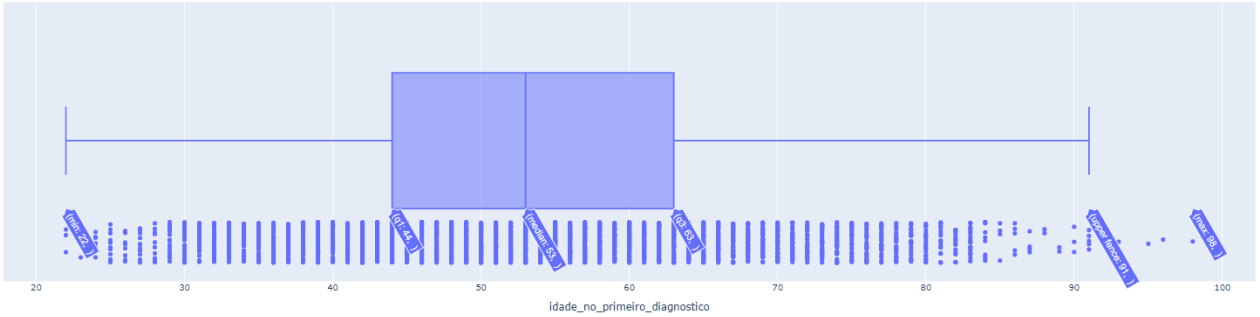


Gráfico número 6:

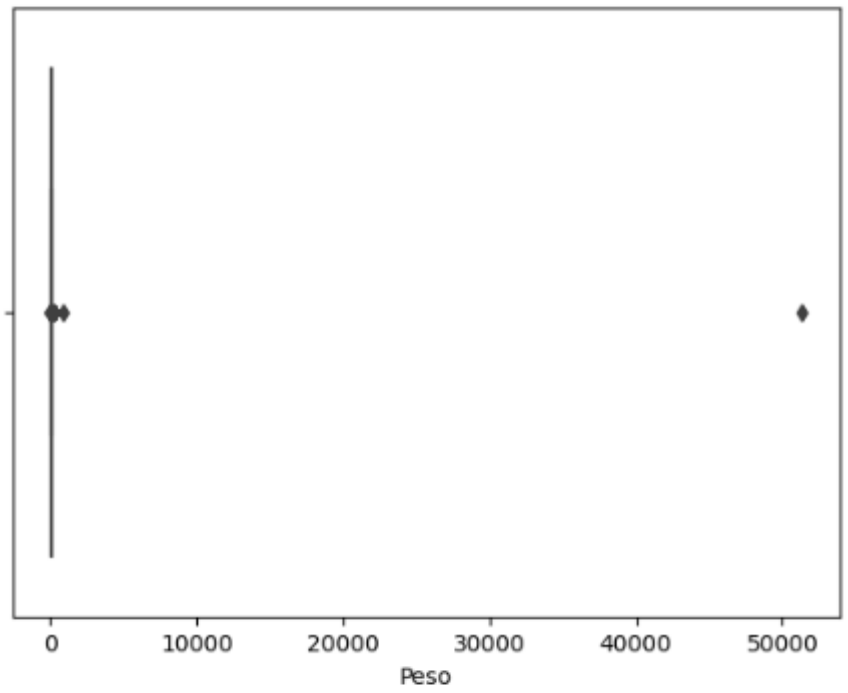
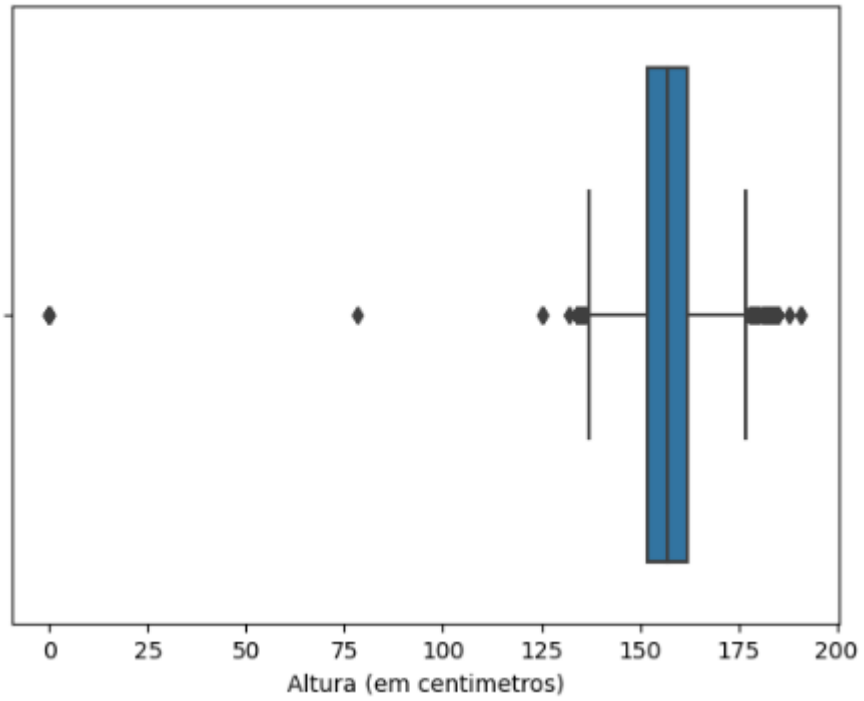


Gráfico número 7:



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:

a) 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:

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 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.

```
# 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
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:

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:

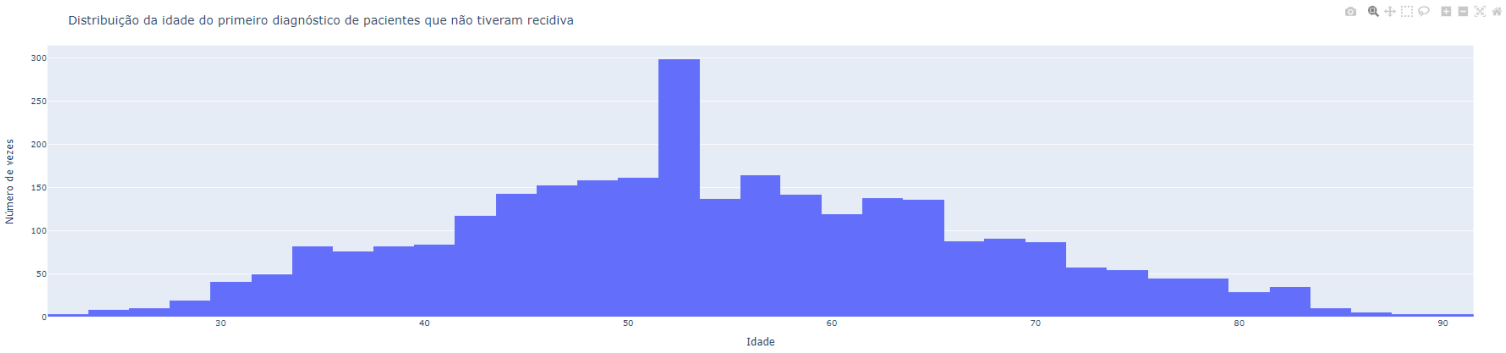
1. Pergunta motivadora: Existe uma relação direta entre o IMC e a gravidade do tumor que influencia na escolha do tratamento sem aliar com a idade?
2. Motivação: Estudos (https://repositorio.uniceub.br/jspui/handle/prefix/13527) (http://repositorio.aee.edu.br/handle/aee/19122)
   Através de pesquisas 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:

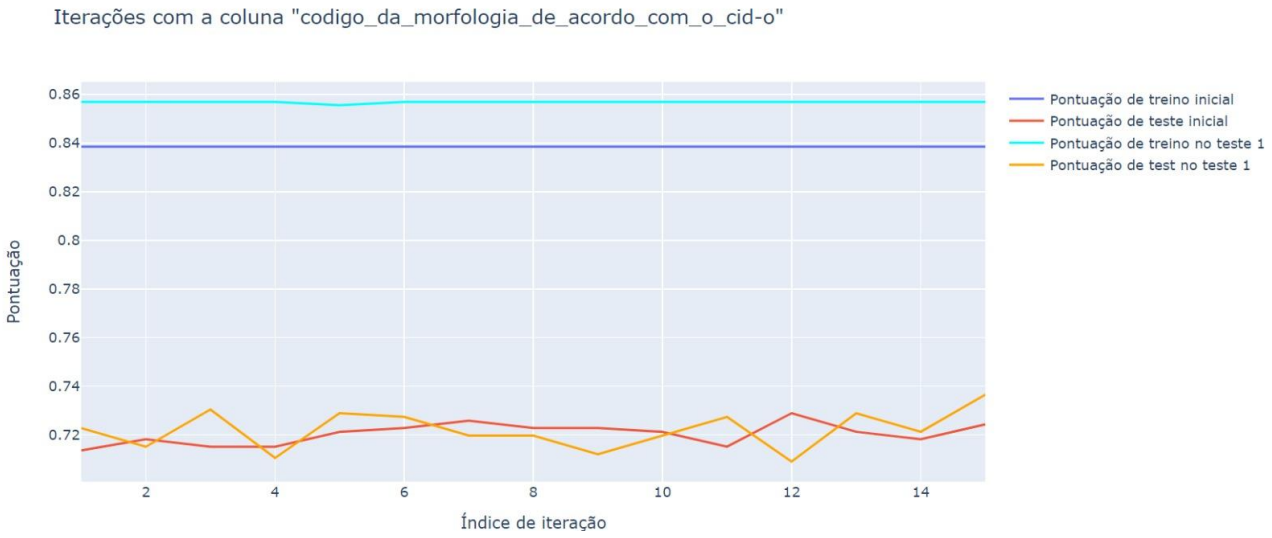
nº	Pontuação treino	Pontuação teste
1	0.8384766907419566	0.7136294027565084
2	0.8384766907419566	0.7182235834609495
3	0.8384766907419566	0.7151607963246555
4	0.8384766907419566	0.7151607963246555
5	0.8384766907419566	0.7212863705972435
6	0.8384766907419566	0.7228177641653905
7	0.8384766907419566	0.7258805513016845
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**



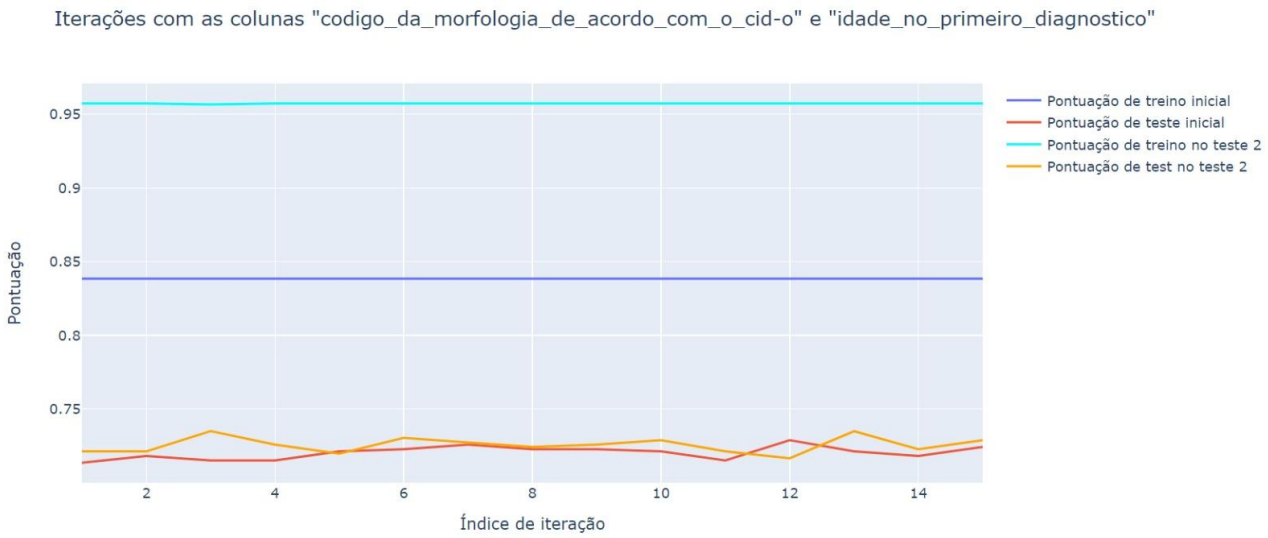
3. Conclusões: Os testes demonstram que existe sim uma relação, mesmo que aparentemente pequena, o código da morfologia acaba influenciando na eficácia do modelo, principalmente nos treinos. Além disso, por ter uma margem pequena de testes, é inconclusivo o impacto que o código da morfologia tem sobre a escolha do tratamento, mas presume-se que ele seja mais positivo do que negativo.
4. 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**



3. 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**

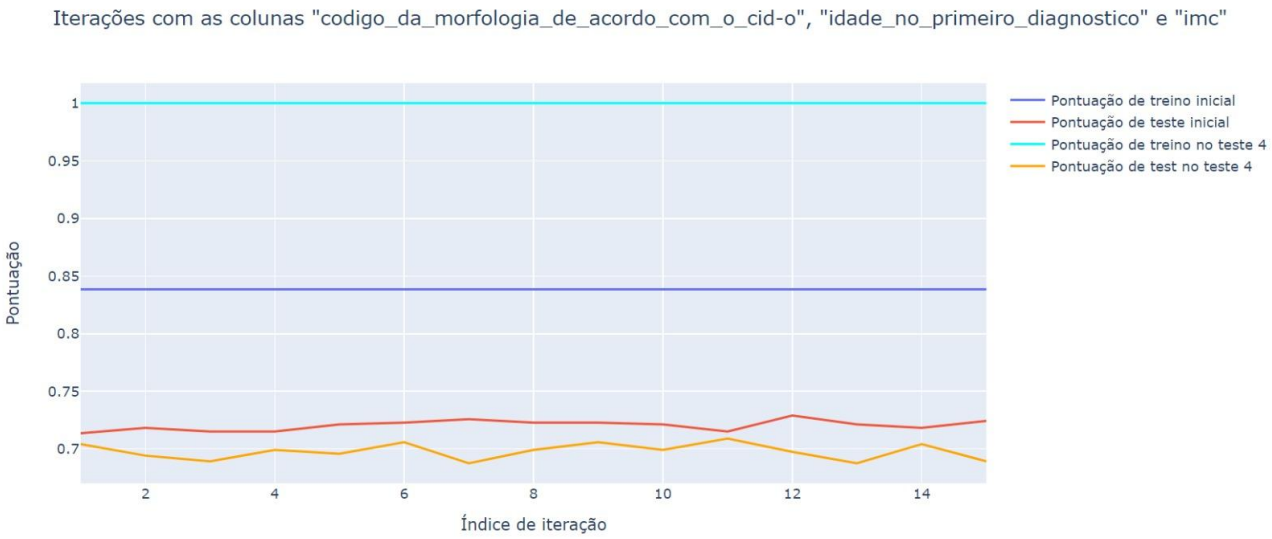
3. Conclusões: Apesar de aparecer um leve declínio na taxa de acertos quando é colocada a variável IMC, as resoluções ainda são inconclusivas, uma vez que estudos demonstram a relação entre tumor e IMC. Existe a necessidade de avaliação futura se o IMC aliado com o fator idade pode ser um indicativo 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:

3. Colunas a serem adicionadas: imc, idade_no_primeiro_diagnostico e codigo_da_morfologia_de_acordo_com_o_cid-o
4. 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**



5. 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://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://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://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.