



OncoAI
Faculdade de
Medicina da USP



Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
02/02/2023	Mariana B. Görresen	1.1	Preenchimento das seções 4.1.3.1, 4.1.3.4
05/02/2023	Mariana B. Görresen	1.2	Preenchimento da seção 4.1.4
05/02/2023	Bruno Wasserstein	1.3	Preenchimento da seção 2
07/02/2023	Mariana B. Görresen	1.4	Preenchimento da seção 4.1.1
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 4.1.2
10/02/2023	Mariana B. Görresen	1.9	Revisão dos tópicos preenchidos
14/02/2023	Bruno Wasserstein e Mariana B. Görresen	2.0	Correção de detalhes e modificação de textos
15/02/2023	Stefano Parente	2.1	Correção da seção 4.1.2
23/02/2023	Bruno Wasserstein e Stefano Parente	2.2	Preenchimento da seção 4.1.8
23/02/2023	Bruno Wasserstein	2.3	Preenchimento da seção 4.2.3 e letra c) da seção 4.2.1
24/02/2023	Mariana Görresen	2.4	Preenchimento da seção 4.2.2 e letras a) e b) da seção 4.2.1
08/03/2023	Stefano Parente	3.0	Preenchimento da seção 3
08/03/2023	Bruno Wasserstein	3.1	Preenchimento da seção 4.3.1 e adequação das referências bibliográficas para as normas da ABNT
10/03/2023	Bruno Wasserstein e Mariana Görresen	3.2	Preenchimento das seções 4.3.2 e 4.3.3
24/03/2023	Mariana Görresen	4.0	Preenchimento da seção 4.4

			Revisão de todos os tópicos.
03/04/2023	Mariana Görresen	5.0	
04/04/2023	Bruno Wasserstein	5.1	Correção de erros e revisão
07/04/2023	Mariana Görresen e Bruno Wasserstein	5.2	Preenchimento da seção 4.5 e 5.0

Sumário

1. Introdução	4
2. Objetivos e Justificativa	5
2.1. Objetivos	5
2.2. Proposta de Solução	5
2.3. Justificativa	5
3. Metodologia	6
4. Desenvolvimento e Resultados	7
4.1. Compreensão do Problema	7
4.1.1. Contexto da indústria	7
4.1.2. Análise SWOT	7
4.1.3. Planejamento Geral da Solução	7
4.1.4. Value Proposition Canvas	7
4.1.5. Matriz de Riscos	7
4.1.6. Personas	8
4.1.7. Jornadas do Usuário	8
4.1.8. LGPD	
4.2. Compreensão dos Dados	9
4.2.1. Exploração dos dados	
4.2.2. Pré processamento	
4.2.3. Hipóteses	
4.3. Preparação dos Dados e Modelagem	10
4.4. Comparação de Modelos	11
4.5. Avaliação	12
5. Conclusões e Recomendações	13
6. Referências	14
Anexos	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, mas 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 a 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 Instituto de Tecnologia e Liderança (Inteli) 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 médicos 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

CRISP-DM é um modelo de processo para projetos de mineração de dados que consiste em seis etapas: Entendimento do negócio, Entendimento dos dados, Preparação dos dados, Modelagem, Avaliação e Implantação. É amplamente utilizado na indústria de análise de dados para maximizar a eficácia de projetos de mineração de dados.

- **Entendimento do negócio:**

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

- **Entendimento dos dados:**

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

- **Preparação dos dados:**

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

- **Modelagem:**

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

- **Avaliação:**

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

- **Implantação:**

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

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

4.1.1.1. Análise da indústria

Ao analisar o contexto do ICESP 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:

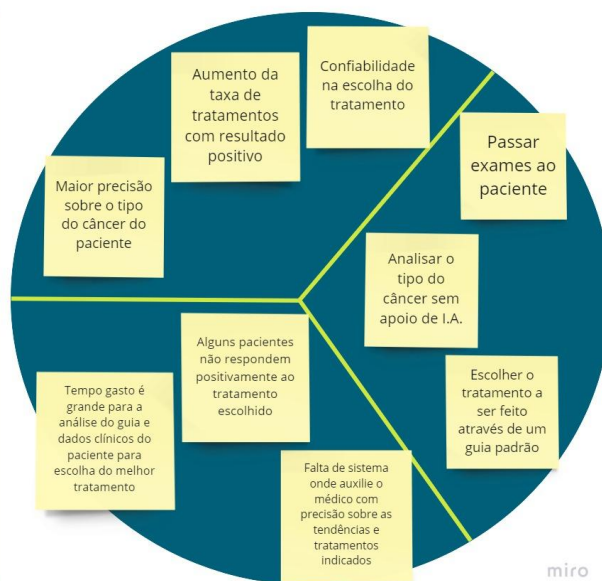
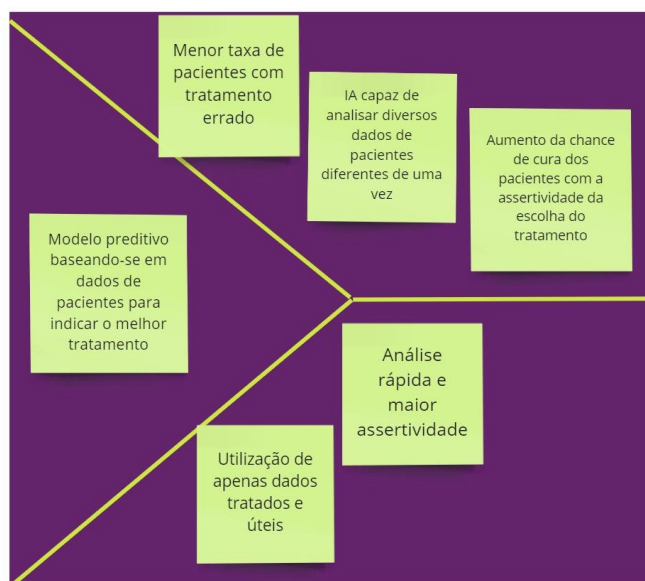
Os principais benefícios são o apoio ao médico no diagnóstico e escolha do tratamento, esse apoio será obtido por meio de auxílios vindos da IA sem tirar o lugar do médico e, sim, otimizando seu trabalho. Através desse auxílio, o médico terá maior chance de ter indicado o melhor tratamento para aquele 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.

Novo: O critério de sucesso será obter um índice de acerto na recomendação do tratamento mais adequado em pelo menos 75% de acurácia. Com o treinamento, escolha adequada de hiperparâmetros e escolha correta do modelo de machine learning, usaremos a acurácia para determinar se a equipe teve sucesso. Além disso, outro critério de sucesso, o principal, é após o paciente receber o tratamento, ter no mínimo 2 anos sem recidiva.

4.1.4. Value Proposition Canvas



4.1.5. Matriz de Riscos

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


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

4.1.6. Personas

Persona 1 (utiliza o modelo):

PERFIL

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



BIOGRAFIA

Dra. Maria Beatriz, mastologista com 14 anos de experiência em diagnóstico e tratamento de câncer de mama. Ela tem como principal objetivo fornecer o melhor cuidado para seus pacientes. A Dra. está sempre em busca de novas tecnologias e soluções que possam ajudá-la a aprimorar o diagnóstico e o tratamento do câncer de mama. Ela acredita que um modelo preditivo para apoiar na escolha de tratamento para o 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 prognósticos mais precisos
- Confirmar o prognóstico oferecido

MOTIVAÇÕES

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

DORES E FRUSTRAÇÕES

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

Persona 2:

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

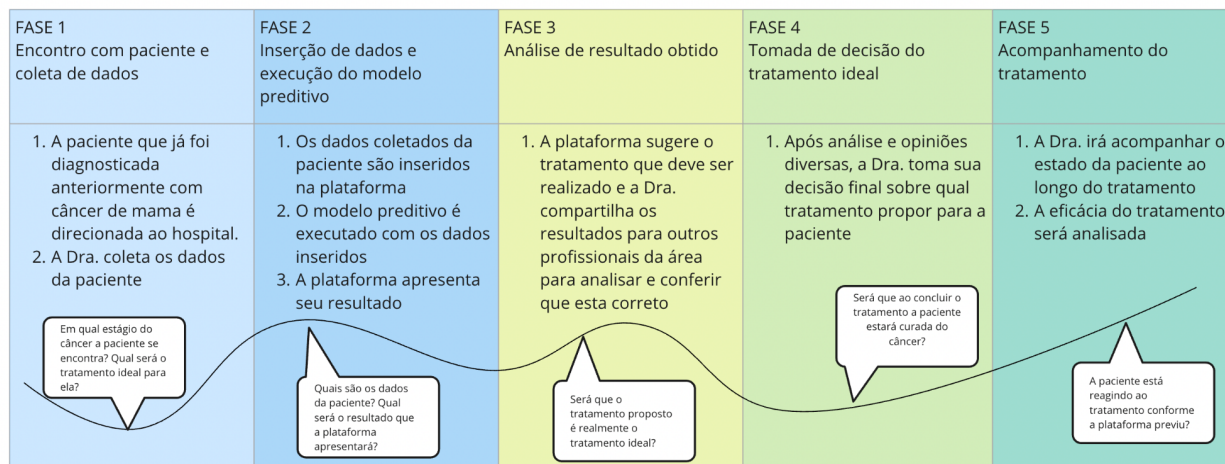


Dra. Maria Beatriz

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

Expectativas

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



Oportunidades

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

Responsabilidades

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

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

Política de Privacidade da OncoAI

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

Coleta de dados:

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

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

Uso dos dados:

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

Armazenamento de dados:

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

Compartilhamento de dados:

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

Segurança de dados:

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

Direitos dos usuários:

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

Atualização da política de privacidade:

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

4.2. Compreensão dos Dados

4.2.1. Exploração de dados:

a) Cite quais são as colunas numéricas e categóricas:

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

Código:

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

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

Código:

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

Output (nome das colunas numéricas):

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

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

Código:

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

Output (nome das colunas categóricas):

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

```

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

```

b) Estatística descritiva das colunas.

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

Método utilizado para fazer a análise descritiva das colunas numéricas:

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

Exemplo do output:

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

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

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

Método utilizado para fazer a análise descritiva das colunas categóricas:

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

Exemplo do output:

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

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

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

c) Gráficos de correlação da estatística descritiva

Gráfico número 1:

Gráfico Idade no primeiro diagnóstico X Atividade física

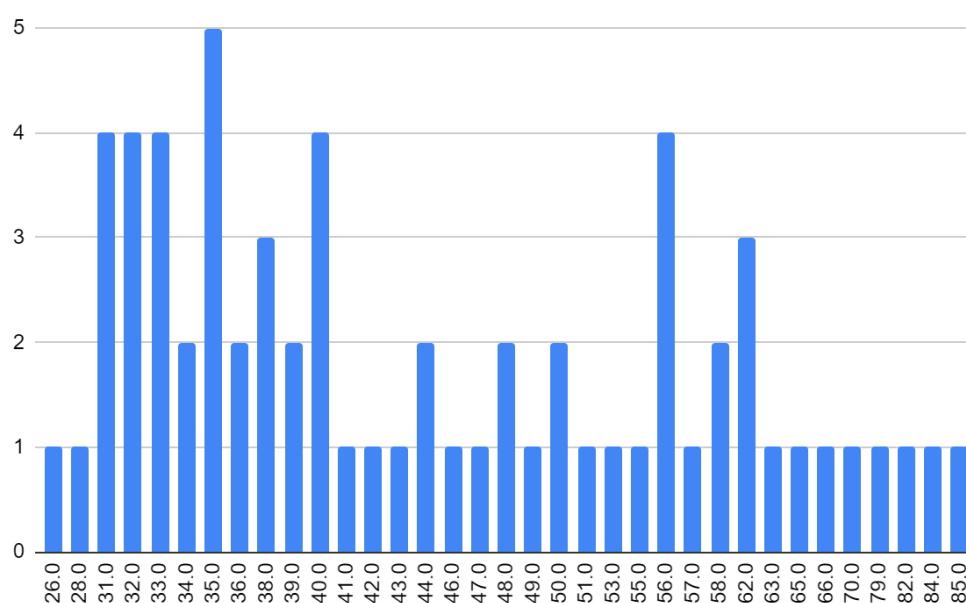
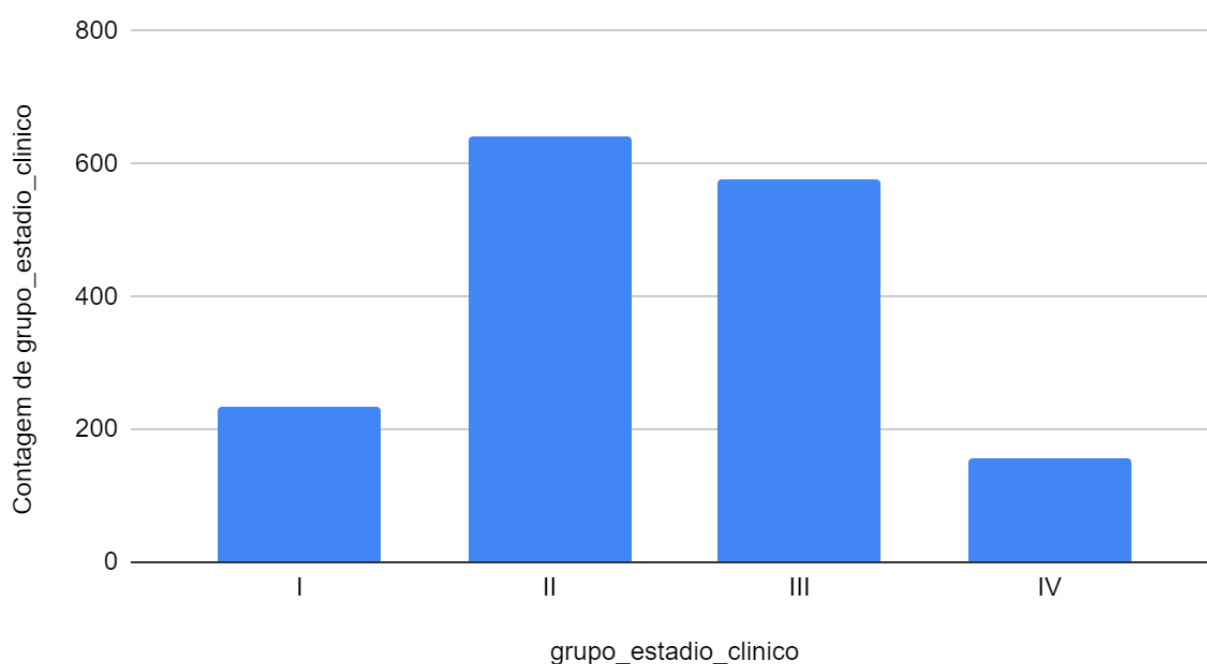


Gráfico que demonstra a relação entre idade e a prática de atividade física (Regular ou frequente) com a idade das pacientes no primeiro diagnóstico, sendo a idade representada no eixo X e a quantidade pacientes que realizam tais atividades no eixo Y.

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

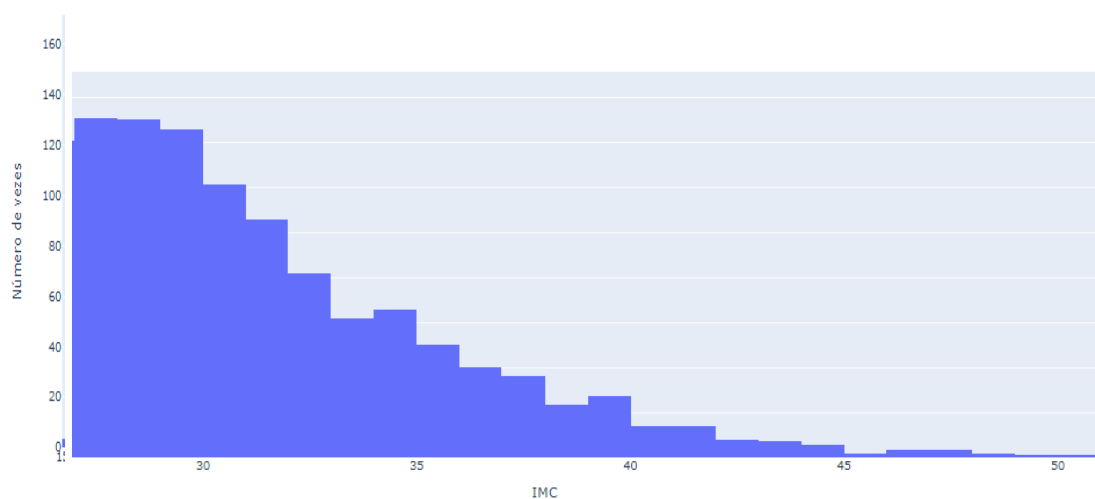
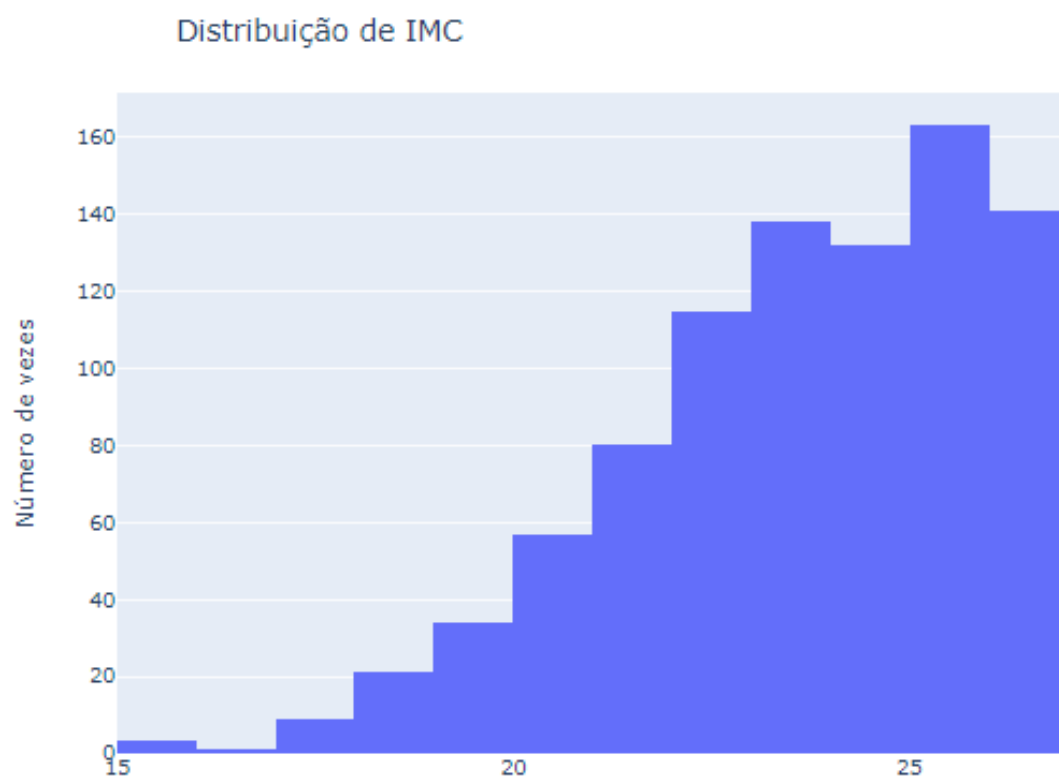
Gráfico número 2:

Contagem de grupo_estadio_clinico



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

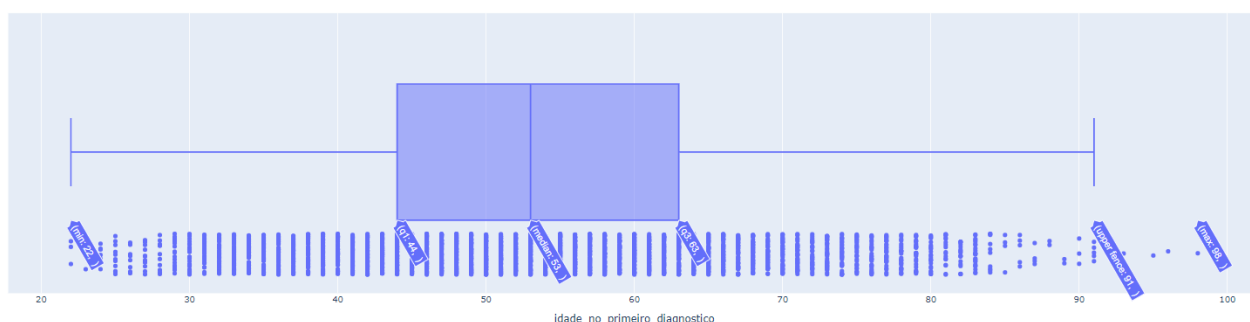
Gráfico número 3:



Conclusão:

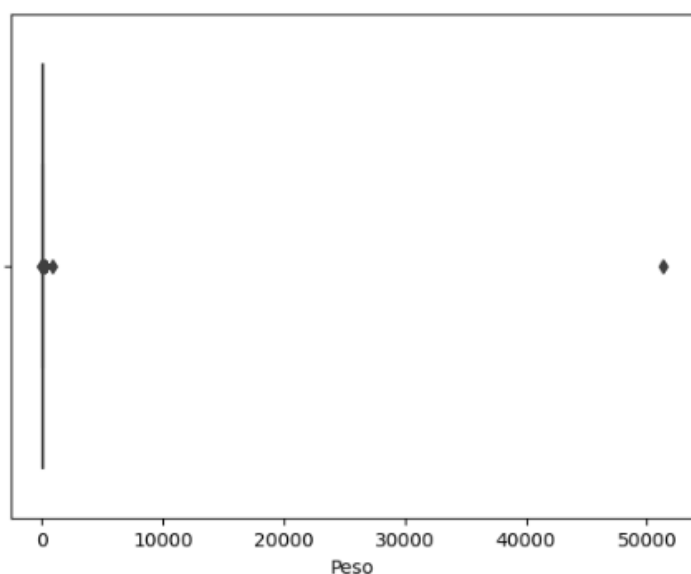
Distribuição de IMC: O gráfico demonstra a distribuição do IMC em pacientes com câncer de mama, é possível perceber uma forte tendência de pacientes com sobrepeso e obesidade, representando uma parcela considerável da quantidade de pacientes. A incidência de sobrepeso e obesidade representa a maioria 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:



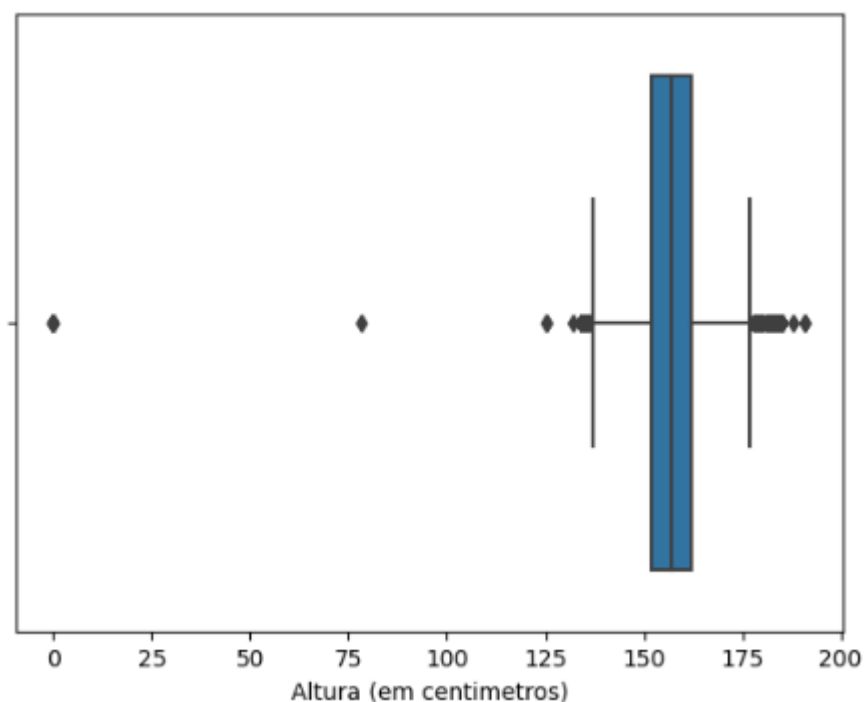
O gráfico número 5 é estilo boxplot. Através dele, podemos analisar os valores da coluna idade_no_primeiro_diagnostico. Havendo uma análise descritiva visual sobre a mesma, assim pode-se identificar os principais pontos de uma coluna para começar a limpeza dos dados, esses são: mediana, q1, q3, limite mais baixo e limite mais alto dos valores, inclusive é possível visualizar possíveis outliers que existem na coluna. Podemos concluir que a idade desses pacientes é maiormente entre 20 a 85 anos, o que podemos supor que idades acima de 85/90 sejam outliers.

Gráfico número 6:



O gráfico número 6 é em formato boxplot. Analisando esse gráfico plotado sobre a coluna Peso, podemos verificar, como no gráfico 5, sua análise descritiva visual e seus valores. É possível identificar, através da análise gráfica, que há grandes outliers na coluna Peso do banco de dados, esses erros, quando muito extrapolantes, são geralmente ocasionados por erros de digitação. Além disso, não conseguimos visualizar a tendência de pesos das pacientes do banco de dados, isso acontece pelos outliers, excedendo demais os valores predominantes, que tornam o gráfico difícil de visualizar.

Gráfico número 7:



O gráfico número 7, também, é um gráfico projetado como boxplot, visando entender como os dados estavam se comportando na coluna Altura (em centímetros). Verificamos que, a altura desses pacientes varia, em maioria, entre 125 e 175 centímetros. Portanto, assim como os gráficos 5 e 6, haviam valores deturpantes que podem ser considerados outliers.

4.2.2. Pré-processamento dos dados:

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

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

Figura 1: Método para encontrar pesos menores que 30.

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

Fonte: Código desenvolvido pelos autores.

Figura 2: Método para encontrar pesos maiores que 150.

```
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

Fonte: Código desenvolvido pelos autores.

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.

Figura 3: Método para encontrar o Record ID do número especificado.

`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

Fonte: Código desenvolvido pelos autores.

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

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

Figura 4: Método para preencher os valores nulos da coluna *Altura (em centímetros)*.

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

Record ID	Altura (em centímetros)
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 centímetros), Length: 3803, dtype: float64

Fonte: Código desenvolvido pelos autores.

Figura 5: Código que junta as colunas que estão corretamente preenchidas.

```
[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 centímetros)_x'], axis=1)
df3 = df3.rename(columns={'Altura (em centímetros)_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

Fonte: Código desenvolvido pelos autores.

Antes e depois de um Record ID que havia outlier:

Figura 5: Demonstração do banco de dados antes do tratamento.

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

	Record ID	Repeat Instrument	Repeat Instance	data_entrada	Peso	Altura (em centímetros)	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

Fonte: Impressão do Google Colab.

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.

Figura 5: Demonstração do banco de dados 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

Fonte: Impressão do Google Colab.

4.2.3. Hipóteses:

Hipótese 1: A morfologia do tumor tem relevância na escolha do tratamento a ser empregado.

1. Motivação: Estudos

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

(http://objdig.ufrj.br/50/teses/m/CCS_M_MarceloSobralLeite.pdf)

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

Hipótese 2: A idade tem relevância na escolha de tratamento.

1. Motivação: estudos

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

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

(http://www.ffclrp.usp.br/imagens_defesas/31_05_2010_17_13_48_43.pdf)

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

Hipótese 3: Existe uma relação direta entre o IMC e o tratamento escolhido.

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

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

Por meio de pesquisas, foi possível perceber que a saúde do paciente está completamente relacionada ao IMC dela. Percebe-se que o valor do IMC pode indicar problemas na saúde do paciente, tanto um IMC com valor muito alto, quanto muito baixo. Por conta do estado de saúde do paciente, com os indícios através do IMC, é provável que o IMC do paciente indique e influencie na escolha do tratamento, por conta dos efeitos colaterais de cada um dos tratamentos.

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

4.3. Preparação dos Dados e Modelagem

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

4.3.1. Modelagem para o problema:

4.3.1.1 Features utilizadas no modelo preditivo:

- Features para realizar os treinos e teste:

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

ki67_maior_14_percentage - O ki67 é um tipo de marcador que ajuda a verificar a taxa de replicação celular das células tumorais. Dessa forma, a literatura médica estabeleceu que um valor em torno de 14% indica um certo grau elevado dessas replicações. Essa medida é um dos critérios para avaliar a gravidade do tumor..

subtipo_tumoral - O subtipo tumoral é uma medida importante de classificação do tumor. De maneira simplificada, essa classificação indica as características do tumor. As classificações são Luminal A, Luminal B, HER2 positivo e Triplo-Negativo. Os tratamentos podem variar conforme o subtipo tumoral e por isso é uma feature relevante e precisa ser considerada para prever o tratamento correto.

receptor_de_progesterona - O receptor hormonal ajuda a indicar o tipo de tratamento para o câncer de mama. Mais especificamente, ajuda a indicar se é necessário aliar o tratamento com hormonoterapia. Porém, além do receptor de progesterona positivo, é necessário ser constatado receptor de estrogênio positivo, para indicação de hormoterapia.

receptor_de_estrogenio - O receptor de estrogênio indica, sozinho ou aliado ao receptor de progesterona, se o tratamento necessitará de hormonoterapia e qual o tratamento mais adequado.

her2_por_ihc - O IHC é um método para detecção da proteína HER2. Quando essa proteína é detectada, significa que o tumor não é do tipo mais comum e que o tratamento para esse tumor precisa ser feito seguindo o padrão anti-HER2. Dessa forma, indica que o tratamento difere do tratamento de câncer HER2 negativo.

her2_por_fish - O fish é um exame para detecção da proteína HER2. Ele ajuda a indicar se existe a presença dela e no caso, como consequência, ajuda na escolha do tratamento adequado.

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

classificacao_tnm_clinico_t - A classificação TNM-T indica o tamanho do tumor, impactando diretamente na escolha de terapia adjuvante ou neoadjuvante, pois quando é necessário reduzir o tamanho do tumor, opta-se pela neoadjuvante.

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

- Target do modelo preditivo:

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

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

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

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

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

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

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

4.3.2. Métricas relacionadas ao modelo:

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

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

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

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

Figura 2: Demonstração do cálculo da acurácia.

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

Fonte: imasters.com.br

Já o F1-score é uma média harmônica das métricas Precision e Recall. Para contextualizar, a métrica Precision mostra apenas o valor dos positivos da matriz de confusão, deixando de fora todos os valores considerados negativos. Enquanto, a métrica Recall é exatamente o oposto, ela deixa todos os valores considerados positivos para dentro, o que, em consequência, causa com que alguns valores considerados negativos sejam utilizados. Porém, como o nosso modelo preditivo não é um binário de questão de positivo e negativo, e sim sobre dois tratamentos que podem ser indicados, essas duas métricas, individualmente, não são úteis para ele. Então, por conta dos fatores citados acima, foi escolhida a métrica F1-score, que é uma métrica de avaliação de desempenho de um modelo de classificação, ela considera tanto a precisão quanto a recall do modelo.

Figura 3: Demonstração do cálculo do F1-score.

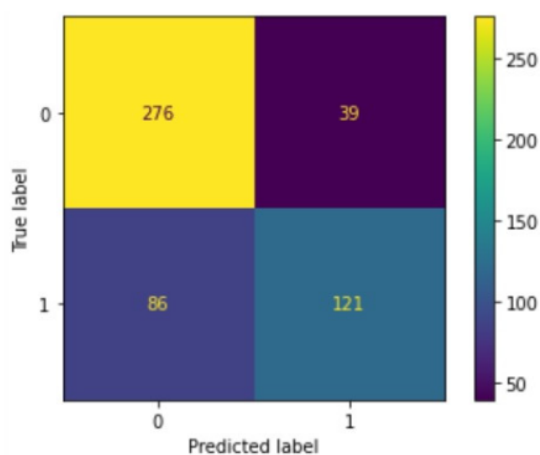
$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fonte: imasters.com.br

Exemplos das métricas nos modelos preditivos supervisionados testados:

Stacking Classifier - Modelo escolhido

Figura 5: Matriz de confusão do modelo Stacking Classifier.



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

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics.

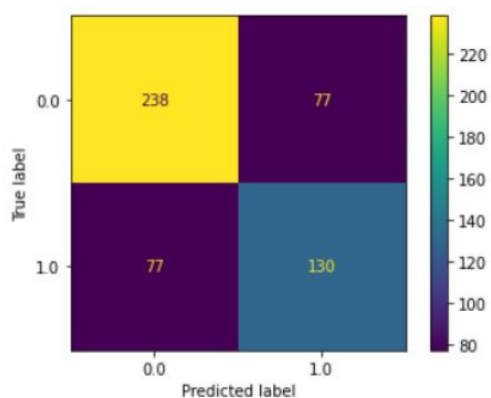
Acurácia Treino: 0.7417763157894737

Acurácia Teste: 0.7605363984674329

F1-score: 0.6594005449591281

Gaussian Naive Bayes - Modelo testado

Figura 6: Matriz de confusão do modelo Gaussian Naive Bayes:



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

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics

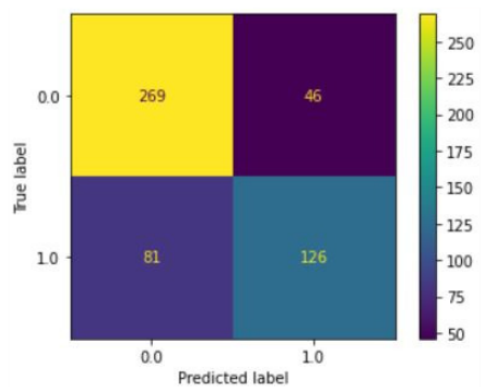
Acurácia Treino: 0.7302631578947368

Acurácia Teste: 0.7049808429118773

F1-score: 0.6280193236714976

Logistic Regression - Modelo testado

Figura 7: Matriz de confusão do modelo Logistic Regression:



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

Fonte: Biblioteca matplotlib.pyplot e sklearn.metrics

Acurácia Treino: 0.7368421052631579

Acurácia Teste: 0.7567049808429118

F1-score: 0.6649076517150396

4.3.3. Primeiro modelo candidato:

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

Figura 6: Código do modelo preditivo Stacking Classifier.

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

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

n_train = []
n_test = []

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

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

y_pred1 = alg.predict(X_test)

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

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Primeiramente, olhando a acurácia, pode-se perceber que não houve overfitting - quando os dados se especializam demais no conjunto de teste com uma diferença maior que 0.15 entre os conjuntos de treino e teste. Também não houve underfitting - quando os resultados dos conjuntos de treino e teste tem valores muito abaixo do esperado - concluindo que ele não ficou muito enviesado nos treinos, o que é considerado um ponto positivo por demonstrar grande margem para melhora da acurácia sem comprometer os treinos. Esse modelo preditivo obteve resultados de testes com maior acurácia em comparação aos outros, foi o único modelo que chegou em 0.76 de acurácia no conjunto de testes, enquanto outros modelos testados ficaram entre 0.70 e 0.75.

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

Sobre a matriz de confusão do Stacking Classifier, pode-se observar que os valores verdadeiros da diagonal principal da matriz de confusão foram os com maioria de resultados, ou seja, o modelo está acertando o tratamento na maioria das vezes. Porém, pode-se observar que o modelo está acertando na previsão de terapia adjuvante de forma consideravelmente maior do

que na terapia Neoadjuvante, onde há maior número de erros. Sendo assim, podemos concluir que ainda é necessário regular melhor o modelo e explorar mais possibilidades para conseguir reduzir os falsos resultados de neoadjuvantes e adjuvantes que o modelo gera, e entender porque há uma preferência do modelo pelo tipo de tratamento adjuvante, uma vez que estes não são os resultados desejados, já que queremos ter resultados onde tenha uma porcentagem correta balanceada para os dois tipos de tratamento.

4.4. Comparação de Modelos

4.4.1. Escolha da métrica e justificativa:

A métrica escolhida foi a acurácia. De forma simplificada, a acurácia mede a proporção de predições corretas feitas pelo modelo em relação ao total de predições realizadas. Essa métrica é facilmente compreensível, até mesmo por pessoas menos familiarizadas com métricas, o que a torna uma opção atraente em diversos contextos. Outra vantagem da acurácia é que ela permite uma medida fácil de relacionar entre efetividade do resultado e valor máximo possível. Embora o valor máximo possa ser difícil ou mesmo impossível de alcançar na prática, a acurácia estabelece um limite claro que torna mais fácil comparar o desempenho de diferentes modelos ou algoritmos.

4.4.2. Modelos otimizados:

Os modelos candidatos e otimizados para o problema em questão foram: AdaBoost Classifier, Logistic Regression e Stacking Classifier. Esses algoritmos foram selecionados com base em diversos testes realizados onde foram analisados os melhores resultados em relação à métrica de acurácia.

4.4.2.1 AdaBoost Classifier:

O Adaboost Classifier é um algoritmo que cria um modelo a partir de vários modelos fracos. Ele treina inicialmente um modelo fraco com pesos iguais para todos os dados e, em seguida, atribui pesos maiores aos dados classificados incorretamente. Esse processo é repetido diversas vezes, treinando modelos fracos até que receba um com peso menor aos erros. Resultando, no final, em um modelo forte feito por vários modelos fracos.

Figura 7: Algoritmo de AdaBoost Classifier sem tuning de hiperparâmetros.

```
[ ] adaboost = AdaBoostClassifier()

    adaboost.fit(X_train , y_train)

    y_pred = adaboost.predict(X_test)
    conf_matrix = confusion_matrix(y_test, y_pred)
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 8: Resultado das métricas de acurácia, dos conjuntos treino e teste, sem hiperparâmetros.

```
Acuracia de treino: 0.7984126984126985
Acuracia de teste: 0.7814726840855107
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Podemos concluir que o modelo não está sofrendo overfitting - quando o modelo se especializa demais no conjunto de treino e devolve um resultado ruim no conjunto de teste - e, também, não está sofrendo underfitting - quando o resultado dos conjuntos de treino e teste não mostram bons resultados. Além disso, estão em números consideravelmente altos, ou seja, bons resultados. Isso significa que o modelo AdaBoost Classifier sem otimização de hiperparâmetros está fazendo previsões boas para o problema em questão.

Figura 9: Algoritmo de AdaBoost Classifier com tuning de hiperparâmetros.

```
[ ] modelo = AdaBoostClassifier()

# Cria o GridSearchCV

parameters = {'algorithm': ['SAMME.R', 'SAMME'],
              'learning_rate': [1.0, 2.0, 3.0],
              'n_estimators': [50, 100, 200],
              'random_state': [500, 501, 502, 503]}

modelGS = GridSearchCV(modelo, parameters, cv=5, scoring='accuracy')

# Treina os modelos e guarda na variável modelGS o melhor modelo
modelGS.fit(X_train, y_train)
modelGS.best_params_
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 10: Melhores parâmetros encontrados pelo Grid Search.

```
{'algorithm': 'SAMME',
 'learning_rate': 1.0,
 'n_estimators': 100,
 'random_state': 500}
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Para fazer a otimização de hiperparâmetros no modelo de aprendizado de máquina AdaBoost Classifier, utilizamos o algoritmo Grid Search. O Grid Search trabalha intensivamente, verificando cada combinação possível e existente na grade de valores da variável "parameters". Ou seja, o seu objetivo é, de fato, encontrar o melhor desempenho do modelo, resultando na melhor combinação de hiperparâmetros para o modelo.

Figura 11: Resultado das métricas de acurácia, dos conjuntos treino e teste, com hiperparâmetros.

```
Acurácia treino: 0.7793650793650794  
Acurácia teste: 0.7909738717339667
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Pode-se notar, após a otimização de hiperparâmetros, que a acurácia do treino diminuiu levemente enquanto a acurácia de teste aumentou. Portanto, pode indicar que a otimização de hiperparâmetros resultou em uma melhor capacidade de generalização do modelo.

4.4.2.2 Logistic Regression:

O Logistic Regression é um algoritmo de aprendizado supervisionado de classificação que utiliza matemática para encontrar relação entre duas variáveis para fazer previsões. Como é um algoritmo de classificação, a regressão logística é utilizada prevendo a probabilidade de um registro pertencer a uma determinada classe com base nas variáveis de entrada.

Figura 12: Algoritmo de Logistic Regression sem tuning de hiperparâmetros.

```
logic = LogisticRegression()  
logic.fit(X_train , y_train)  
  
y_pred = logic.predict(X_test)
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 13: Resultado das métricas de acurácia, dos conjuntos treino e teste, sem hiperparâmetros.

```
Acuracia de treino: 0.7603174603174603  
Acuracia de teste: 0.7957244655581948
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Analisando a acurácia do modelo, podemos perceber que a acurácia do conjunto de teste é consideravelmente maior que a de treino, mas não o suficiente para serem considerados resultados ruins. Além disso, o resultado do conjunto de teste está levemente maior que o resultado do conjunto de teste do algoritmo AdaBoost Classifier, indicando que o modelo Logistic Regression, sem hiperparâmetros, se saiu levemente melhor do que a versão sem hiperparâmetros do AdaBoost Classifier.

Figura 14: Algoritmo de Logistic Regression com tuning de hiperparâmetros.

```
modelo = LogisticRegression()

# Cria o RandomSearchCV

parameters = {'C': [0.01, 0.1, 1],
              'intercept_scaling': [1, 2, 3],
              'l1_ratio': [None, 0.1],
              'max_iter': [10, 100, 1000],
              'multi_class': ['auto', 'ovr', 'multinomial'],
              'penalty': ['l1', 'l2', 'elasticnet', None],
              'random_state': [1, 100, 6664],
              'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
              'tol': [0.1, 0.001, 0.0001],
              'verbose': [0, 1, 2]}

modelRS = RandomizedSearchCV(modelo, parameters, cv=5, scoring='accuracy')

# Treina os modelos e guarda na variável modelRS o melhor modelo
modelRS.fit(X_train, y_train)
modelRS.best_params_
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 15: Melhores parâmetros encontrados pelo Randomized Search.

```
'tol': 0.001,
'solver': 'newton-cg',
'random_state': 6664,
'penalty': None,
'multi_class': 'ovr',
'max_iter': 10,
'l1_ratio': 0.1,
'intercept_scaling': 2,
'C': 0.01}
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Escolhemos o algoritmo de otimização de hiperparâmetros Randomized Search para aplicar no modelo de aprendizado de máquina Logistic Regression. Ao contrário do Grid Search, onde o algoritmo experimenta cada combinação da grade de parâmetros para obter o melhor desempenho, o Randomized Search não funciona de forma tão intensa computacionalmente. Ele utiliza de uma técnica mais simples e prática, escolhendo de forma aleatória e pulando possíveis combinações, onde o algoritmo pode rodar de forma mais rápida e sem sobrecarregar o servidor. A escolha do Randomized Search, para o modelo Logistic Regression, se deu pelo fato de que esse aprendizado de máquina obtém um número maior de parâmetros para implementar na otimização de hiperparâmetros, o que causaria, se fosse utilizado o Grid Search, uma sobrecarga e demora extrema da realização do código para conseguir a melhor combinação possível.

Figura 16: Resultado das métricas de acurácia, dos conjuntos treino e teste, com hiperparâmetros.

```
Acc treino: 0.7571428571428571
Acc teste: 0.7957244655581948
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Através da comparação e análise dos resultados de acurácia, antes e depois dos hiperparâmetros, podemos perceber que a acurácia de treino diminuiu após o tuning de hiperparâmetros e a de teste não obteve nenhuma melhora ou piora. Sendo assim, a conclusão que podemos obter é que, após a otimização, o desempenho do modelo piorou e mostrou não ter uma boa generalização. Portanto, em comparação com o modelo AdaBoost Classifier, tanto otimizado com hiperparâmetros quanto a versão não otimizada, o modelo Logistic Regression não é a melhor opção para o problema em questão, por não apresentar uma melhor generalização de dados.

4.4.2.3 Stacking Classifier:

O modelo de ensino supervisionado, Stacking Classifier, é uma técnica de aprendizado de máquina onde combina múltiplos estimadores, modelos de classificação, para encontrar o melhor desempenho geral. Ele treina vários modelos diferentes classificatórios e usa suas previsões como entrada para fazer a predição final. A vantagem do Stacking Classifier é que ele consegue achar padrões diversos de vários modelos, o que não seria possível utilizando um modelo único de classificação.

Para desenvolver nosso modelo Stacking Classifier, utilizamos como estimadores os modelos de classificação Logistic Regression, SVC e AdaBoost Classifier. O motivo da escolha desses modelos foram diversas pesquisas para entender quais os modelos mais apropriados para nosso problema em questão, e muitos testes para avaliar esses modelos e suas acurácias. O Logistic Regression foi utilizado tanto no conjunto de múltiplos estimadores para fazer previsões de entrada quanto para fazer a predição final, ou seja, ele foi o estimador responsável pela previsão final do modelo.

Figura 17: Algoritmo de Stacking Classifier sem tuning de hiperparâmetros.

```
estimators = [('lr', LogisticRegression()), ('svr', SVC(max_iter=2000)), ('ada', AdaBoostClassifier())]

n_train = []
n_test = []

alg = StackingClassifier(estimators=estimators, final_estimator=LogisticRegression())

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

y_pred1 = alg.predict(X_test)
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 18: Resultado das métricas de acurácia dos conjuntos treino e teste, respectivamente, sem hiperparâmetros.

```
Acurácia treino: [0.7587301587301587]
Acurácia teste:  [0.7838479809976246]
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Fazendo a análise das acurácias do modelo, podemos perceber que há uma diferença entre os valores das acurácias de treino e teste, onde a de teste é melhor. Os resultados não demonstraram valores considerados ruins, dado que estão bem parecidos com os modelos mostrados anteriormente. Porém, em comparação com os resultados dos modelos AdaBoost Classifier e Logistic Regression, nas suas versões não otimizadas, o modelo Stacking Classifier obteve uma pontuação de acurácias menor, indicando que esse algoritmo sem tuning de hiperparâmetros não é a melhor escolha de modelo para o problema.

Para fazer o tuning de hiperparâmetros nesse algoritmo, é necessário fazer a otimização em cada modelo/estimador escolhido. Então, fizemos a otimização de hiperparâmetros dos modelos Logistic Regression, SVC e AdaBoost Classifier - os processos de otimização dos modelos AdaBoost Classifier e Logistic Regression podem ser encontrados nos tópicos [4.4.2.1](#) e [4.4.2.2](#), respectivamente.

Figura 19: Algoritmo de otimização do modelo SVC, com Randomized Search.

```
modelo = SVC()

# Cria o RandomSearchCV

parameters = {'C': [1, 3, 5,],
              'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
              'degree': [3, 5, 7],
              'gamma': ['scale', 'auto'],
              'tol': [1, 0.01, 0.001],
              'max_iter': [-1, 100, 1000],
              'decision_function_shape': ['ovo', 'ovr']}

modelRS = RandomizedSearchCV(modelo, parameters, cv=5, scoring='accuracy')
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 20: Melhores parâmetros encontrados pelo Randomized Search.

```
{'tol': 0.001,
 'max_iter': -1,
 'kernel': 'linear',
 'gamma': 'auto',
 'degree': 7,
 'decision_function_shape': 'ovr',
 'C': 5}
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Figura 21: Algoritmo de Stacking Classifier com hiperparâmetros otimizados.

```
estimators = [('lr', LogisticRegression(C=0.01, intercept_scaling=3, max_iter=1000,
multi_class='ovr', penalty=None, random_state=100,
solver='sag', verbose=1)), ('svr', SVC(max_iter= 2000, C=5, degree=7, gamma='auto', kernel='linear'))
, ('ada', AdaBoostClassifier(algorithm='SAMME', n_estimators=100, random_state=500))]

n_train = []
n_test = []

alg = StackingClassifier(estimators=estimators, final_estimator=LogisticRegression(C=0.01, intercept_scaling=3, max_iter=1000,
multi_class='ovr', penalty=None, random_state=100,
solver='sag', verbose=1))

alg.fit(X_train, y_train)
```

Fonte: Desenvolvido pelos autores através do Google Collaboratory.

Figura 22: Resultado das métricas de acurácia dos conjuntos treino e teste, respectivamente, após otimização de hiperparâmetros.

```
Acurácia treino: [0.7904761904761904]
Acurácia teste: [0.8052256532066508]
```

Fonte: Impressão de tela, através Google Collaboratory, do resultado do código.

Após a análise dos resultados de acurácia do Stacking Classifier otimizado, é perceptível uma melhora na acurácia do modelo, obtendo uma acurácia de 80% nos testes. Além disso, o modelo não demonstra muita diferença nos conjuntos de treino e teste, indicando haver uma melhora na generalização dos dados, considerando, também, os valores obtidos.

4.4.3. Definição do modelo candidato:

Depois dos testes de vários modelos utilizando Grid Search e Random Search, foi possível identificar, através das métricas, que o Stacking Classifier, otimizado com hiperparâmetros, se saiu melhor que os outros modelos. Esse modelo apresentou uma melhor generalização, demonstrou uma acurácia no conjunto de teste melhor e seus conjuntos de treino e teste são balanceados - não há underfitting nem overfitting. Sendo assim, definimos, com base nesses motivos, o Stacking Classifier como modelo de aprendizado de máquina para o problema em questão.

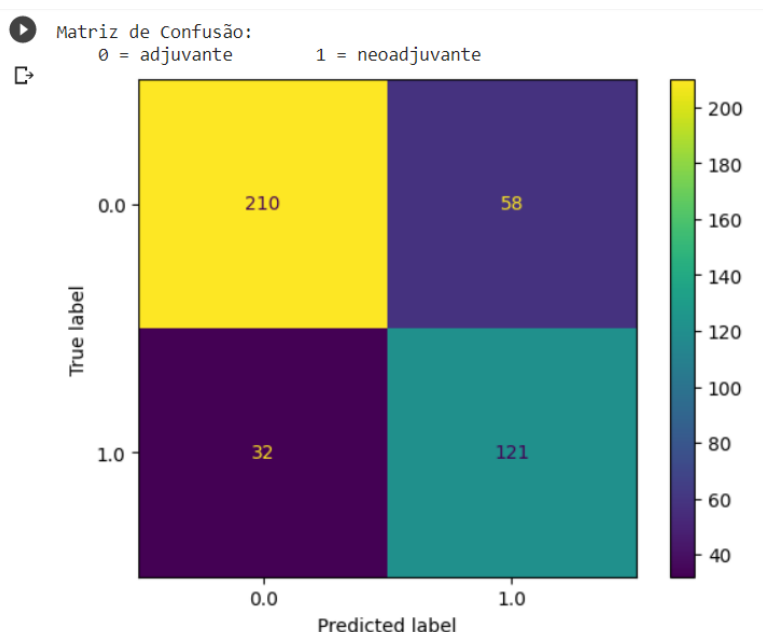
4.5. Avaliação

Nosso modelo final é chamado de CureMate e tem como objetivo auxiliar o médico na escolha da melhor forma de tratamento para o câncer de mama: terapia neoadjuvante - quimioterapia antes da cirurgia - ou terapia adjuvante - cirurgia seguida de quimioterapia. Essa solução atende aos requisitos de alívio de dores e ganhos dos criadores, conforme sinalizado na Seção 4.1.4, e é capaz de entregar o produto/serviço projetado na Proposta de Valor Canvas.

O CureMate fornecerá o resultado da indicação de tratamento por meio de uma aplicação web que utilizará as informações inseridas pelo usuário (médico) para efetuar uma predição com base em padrões encontrados nos dados. Considerando a praticidade para o médico, optamos por disponibilizar a solução final em formato de aplicação web, apresentando-a em um formulário intuitivo e fácil de usar. Entretanto, também há a disponibilização do Google Colab para a utilização do modelo preditivo. Porém, sabemos que disponibilizar apenas em formato de código no Google Colab pode ser desafiador para médicos que não possuem base de programação.

O algoritmo escolhido para a solução final do modelo preditivo foi o Stacking Classifier, que utiliza os modelos estimadores Logistic Regression, Adaboost Classifier e SVC, e obteve uma acurácia final de 77,4%. O conjunto de dados utilizados para treinar o modelo é consideravelmente grande e diversificado, o que permitiu que o modelo encontrasse padrões preditivos e tivesse boa generalização de dados.

Figura 22: Demonstração da matriz de confusão final de predições do modelo.



Fonte: Impressão do Google Colaboratory.

Através da diagonal principal, podemos visualizar a quantidade de predições que o modelo acertou.

No entanto, é importante destacar que o nosso algoritmo de modelo preditivo pode apresentar falhas em suas previsões, pois sua acurácia não é 100%. Por isso, é recomendável que o modelo seja utilizado apenas como mais uma ferramenta para auxiliar na escolha do tratamento e não seja a única maneira de obter o prognóstico do tratamento para o paciente. Em casos de falhas de previsões, é fundamental que sejam utilizados os guias de escolha de tratamento elaborados por médicos capacitados e formados.

É importante que as pessoas entendam a explicabilidade do modelo preditivo, porque assim é possível compreender como o algoritmo do modelo funciona, toma suas decisões e faz suas previsões. Isso é crucial para garantir que o modelo seja ético e impeça de ser uma inteligência artificial enviesada, além de permitir que as pessoas confiem nas previsões que ele gera.

Na fase de pré-processamento e exploração dos dados, os autores procuraram compreender as variáveis dos quatros bancos de dados fornecidos pelo ICESP, em conjunto com a Faculdade de medicina da USP, sobre o histórico de características, fatores de tratamento e sobrevida dos seus pacientes de câncer de mama. Então, no momento inicial foram tratados os dados com erros de digitação, outliers e valores nulos, isso aconteceu para que fosse possível encontrar padrões de correlações entre as variáveis, sem a interferência de dados ruidosos, que indicavam a escolha do tipo de tratamento indicado. Além disso, foram excluídas as colunas onde haviam poucos valores preenchidos ou não eram relevantes para o objetivo do problema em questão.

Após o pré-processamento dos dados, houve a junção das tabelas. Sendo assim, foi feito o *merge* das quatro tabelas com as colunas tratadas para conseguir começar o processo de modelagem dos dados. De início, foi dividido o dataframe em treino e teste, onde 75% do dataframe foi destinado para treinar o modelo preditivo e 25% para ser utilizado como teste.

Figura 22: Código de divisão do dataframe.

```
# Dividindo os dados
X = test.drop(columns=["regime_tratamento"])
y = test["regime_tratamento"]

# Dividindo entre treino e teste para treinar o modelo
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.25,
                                                    random_state = 50)
```

Fonte: Desenvolvido pelos autores.

A partir dessa divisão de dados, foi iniciado o processo de treinar diversos modelos preditivos e escolher as métricas. Cerca de 8 modelos preditivos foram testados, esses são: Random Forest Classifier, Logistic Regression, AdaBoost Classifier, DecisionTree, Gaussian Process Classifier, Bagging Classifier, Linear Regression e Stacking Classifier.

Inicialmente, não foram feitas escolhas de colunas. Então, foram testadas, nos modelos preditivos, todas as colunas e, assim, verificamos algumas hipóteses.

Figura 23: Gráfico de correlações de features, gerado com o Pandas Profiling

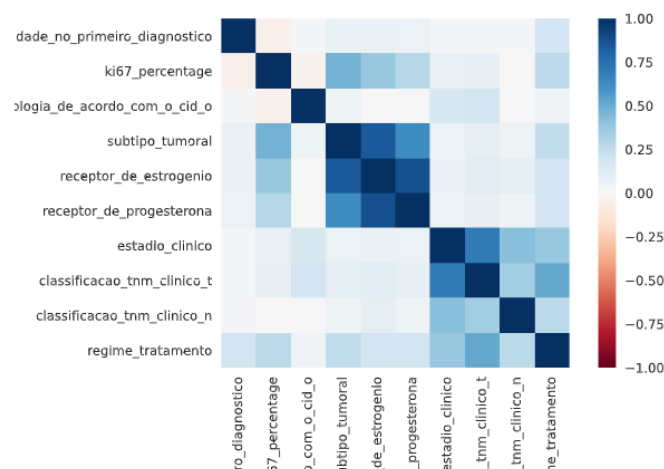
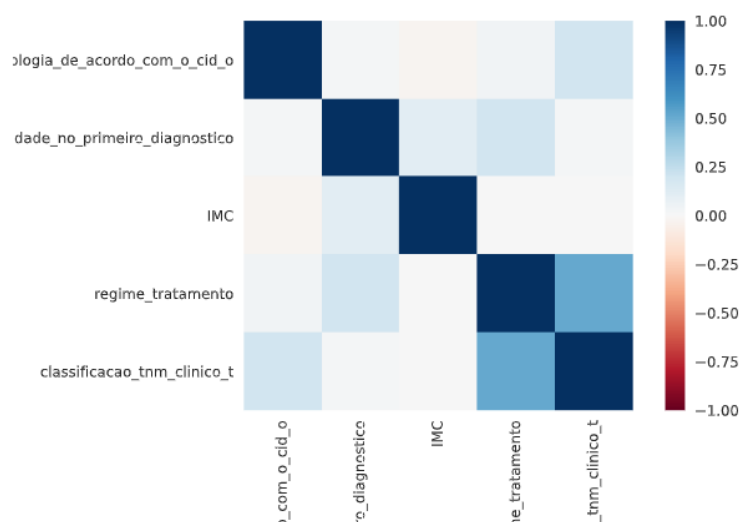


Figura 24: Gráfico de correlações de features, gerado com o Pandas Profiling



A primeira hipótese diz afirma existir uma relação entre a morfologia do tumor e o tratamento escolhido pelo médico para o paciente com câncer. Em sequência, a segunda hipótese fala que existe uma relação entre idade e o tratamento escolhido para o paciente com tumor. E por último, a terceira hipótese diz que o regime de tratamento tem correlação com o IMC do paciente, basicamente se o IMC pode influenciar no tipo de tratamento escolhido.

Utilizando o modelo de machine learning criado pelo time, usando a biblioteca do Pandas Profiling e colocando as features que tem relação com as hipóteses, foi possível concluir, através da visualização da figura 23, que a primeira hipótese foi refutada. O gráfico de correlação demonstra que a morfologia do tumor não teve uma correlação significativa com o regime de tratamento. Pode-se perceber, pelo tom da coloração, que a correlação entre morfologia e regime de tratamento está próximo de 0.

Utilizando o modelo de machine learning, aliado com o Pandas Profiling, ao inserir as features de idade no primeiro diagnóstico e regime de tratamento, para olhar a correlação, foi possível concluir que a segunda hipótese pode ser aceita, apesar da correlação estar num nível entre fraca e moderada, como demonstra na figura 23. Portanto, a segunda hipótese é aceita, por estar próxima de ser uma correlação moderada.

Utilizando o mesmo método para verificação das duas primeiras hipóteses, com a inserção da feature IMC e da feature regime de tratamento, demonstrado na figura 24, é possível perceber que não existe correlação entre o regime de tratamento e o IMC do paciente. Então, podemos refutar a terceira hipótese, que afirmava haver uma relação entre IMC e o regime de tratamento.

Após a análise dos gráficos, conversas com os stakeholders e diversos testes, foram descartadas diversas colunas para treinar o modelo preditivo final, que tem como objetivo encontrar padrões preditivos para descobrir qual o melhor prognóstico do tratamento considerando as características de cada paciente. Por isso, as escolhas foram baseadas em informações obtidas apenas na primeira consulta e tem alguma correlação com a coluna de regime de tratamento.

As colunas escolhidas, para fazer a predição do melhor regime de tratamento indicado, foram: **idade_no_primeiro_diagnostico, subtipo_tumoral, receptor_de_estrogenio, receptor_de_progesterona, ki67_percentage, estadio_clinico, classificacao_tnm_clinico_t, classificacao_tnm_clinico_n.**

Essas colunas foram as que trouxeram melhor acurácia de treino e teste, e melhor acurácia geral do modelo. Além disso, o modelo preditivo final foi treinado apenas com casos de sucesso, onde foram excluídas todas as linhas em que os pacientes tiveram informações finais como morte, vivo com câncer ou recidiva em menos de 5 anos. Assim, o modelo encontra apenas padrões de sucesso e entrega a resposta baseado em casos positivos de um histórico de pacientes variados.

5. Conclusões e Recomendações

Através deste projeto, foi possível constatar de maneira matemática, com resultados precisos de acurácia, quais são os principais fatores determinantes para a escolha de um tipo de tratamento para o paciente. Além disso, é interessante notar que a complementação da base de dados pode melhorar a precisão do modelo, proporcionando assim resultados mais confiáveis e ajudando a tomar decisões mais informadas sobre o tratamento. A integração está sendo realizada entre o modelo programado em Python e uma plataforma web, o que torna o processo de obtenção de resultados mais rápido e conveniente. Para melhorar ainda mais o desempenho do modelo, é necessário continuar a exploração e coleta de dados, principalmente para complementar dados que estão escassos e podem fazer diferença na acurácia.

Uma das melhorias que podem ser implementadas é transformar a aplicação web em algo mais intuitivo e que possa ser utilizado offline, incluindo-a no sistema do hospital. É recomendado que em caso de falhas do modelo preditivo, é de grande importância que os médicos

responsáveis pelo prognóstico continuem a utilizar o guia de escolha de tratamento como sua ferramenta principal de decisão, lembrando sempre que este projeto é apenas uma ferramenta de apoio ao médico e ao paciente no processo de prognóstico do tratamento. Além disso, é importante e recomendado que o médico utilize essa ferramenta junto com o paciente, considerando que é algo que afetará a vida do mesmo, para que haja transparência e base na escolha do seguimento do tratamento.

Na seção Anexos é possível encontrar o link para o Google Colab disponibilizado para uso do usuário.

6. Referências

Agência Internacional de Pesquisa em Câncer (Iarc): O INCA possui assento tanto no Conselho Diretivo como no Conselho Científico. Gov.br. 2022. Disponível em: <https://www.gov.br/inca/pt-br/aceso-a-informacao/institucional/atuacao-internacional/agencia-internacional-de-pesquisa-em-cancer-iarc> Acesso em: 09 mar. 2023.

Modelagem preditiva aumenta eficiência de sistemas de saúde: Com ajuda de inteligência artificial, é possível prever riscos e traçar planos de prevenção em sistemas de saúde. Estadão, 2021. Disponível em: <https://summitsaude.estadao.com.br/saude-humanizada/modelagem-preditiva-aumenta-eficiencia-de-sistemas-de-saude/> Acesso em: 09 mar. 2023.

VIDALE, Giulia. Centro de oncologia do Albert Einstein é eleito o melhor da América Latina: Outros centros brasileiros também integraram a respeitada lista da Newsweek, como o Sírio-Libanês, o A.C. Camargo Câncer Center e a Beneficência Portuguesa. Veja, 2021. Disponível em: Centro de oncologia do Albert Einstein é eleito o melhor da América Latina Leia mais em: <https://veja.abril.com.br/saude/centro-de-oncologia-do-einstein-e-eleito-o-melhor-da-america-latina/> Acesso em: 09 mar. 2023.

Câncer de Mama Receptor de Hormônio. Oncoguia, 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer-de-mama-receptor-de-hormonio/10879/264/> Acesso em: 09 mar. 2023.

Câncer de Mama HER2. Oncoguia, 2017. Disponível em: <http://www.oncoguia.org.br/conteudo/cancer-de-mama-her2/10880/264/>. Acesso em: 09 mar. 2023

ZERWES, Felipe; MILLEN, Eduardo; CAVALCANTE, Francisco Pimentel; NOVITA, Guilherme; FILHO, Hélio Rubens; REIS, João Henrique. Classificações do estadiamento do câncer de mama. câncer de mama Brasil, 2018. Disponível em: <https://www.cancerdemamabrasil.com.br/estadiamento-do-cancer-de-mama/> Acesso em: 09 mar. 2023.

SATO, Rafael Onuki. Você sabe o que é uma metástase?. Dr. Rafael Onuki Sato, 2018. Disponível em: <https://drrafaelsato.com.br/metastase/#:~:text=Qual%20a%20diferen%C3%A7a%20entre%20recidiva,independente%20do%20local%20de%20aparecimento> Acesso em: 09 mar. 2023.

TOLENTINO, Grassyara Pinho. Avaliação da composição corporal, qualidade de vida e toxicidade do tratamento quimioterápico em mulheres com câncer de mama. UnB, 2017. Disponível em: <https://repositorio.unb.br/handle/10482/22356> Acesso em: 09 mar. 2023.

Anexos

Link do Colab do usuário:

https://colab.research.google.com/drive/1fRn2_zcrLFB9_GHTDRMqsxmfs4oSYvnE?usp=sharing