NeoVision
USP Medicina

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
31/01/2023	Marcelo Saadi	1.1	Criação do documento e contexto da indústria
01/02/2023	Vinicius Kumagai Vitor Hugo Rodrigues	1.2	Objetivos e introdução
02/02/2023	Celine Souza Tony Sousa	1.3	Personas e jornadas do usuário
04/02/2023	Guilherme Moura	1.4	Análise SWOT
05/02/2023	José V. Alencar	1.5	Value Proposition Canvas e Matriz de Riscos
05/02/2023	Tony Jonas	1.6	User Stories
06/02/2023	Marcelo Saadi	1.7	Atualização do contexto da indústria
08/02/2023	Tony Jonas	1.8	Correções no Contexto da Indústria, Persona e Jornada do Usuário
10/02/2023	Tony Jonas José Vitor Alencar	1.9	Correções gerais para entrega

Sumário

1. Introdução

O parceiro de negócios do projeto é formado por duas instituições: a Faculdade de Medicina da Universidade de São Paulo e o Instituto de Câncer do Estado de São Paulo.

A Faculdade de Medicina da Universidade de São Paulo (USP) é a principal faculdade de medicina do Brasil, localizada na cidade de São Paulo. Possui o mais alto porte entre as faculdades de medicina do país, sendo reconhecida como uma das melhores do mundo. A faculdade oferece uma grande variedade de programas de estudos em diferentes áreas médicas, incluindo medicina, cirurgia, enfermagem, farmácia, nutrição e, claro, oncologia. Além disso, a USP possui uma forte presença no campo da pesquisa médica, com inúmeros projetos de pesquisa sendo realizados na faculdade.

O Instituto do Câncer do Estado de São Paulo (ICESP) é uma instituição brasileira de referência em tratamento, pesquisa e ensino no âmbito da oncologia. Localizada na cidade de São Paulo, o ICESP possui unidades de atendimento hospitalar e ambulatorial, além de unidades de ensino médico. A instituição tem como objetivo fornecer serviços de excelência a todas as pessoas com câncer, desde o diagnóstico até o tratamento. O ICESP também atua na pesquisa de novas técnicas de tratamento e medicamentos, além de promover a educação médica. A instituição tem se destacado como uma referência para a oncologia no Brasil, sendo reconhecida por seu alto padrão de atendimento e excelência na pesquisa.

Nesse contexto, o problema trazido pelo parceiro é determinar, a partir de dados clínicos dos pacientes, qual o melhor tratamento para o câncer de mama: neo (1º quimioterapia e 2º cirurgia) ou adjuvante (1º cirurgia e 2º terapia).

2. Objetivos e Justificativa

2.1. Objetivos

O principal objetivo do projeto é criar um modelo preditivo que categorize qual tratamento é mais recomendado para casos de câncer de mama para pacientes do Instituto de Câncer de São Paulo (ICESP), de acordo com o perfil e dados disponibilizados desses pacientes. Os tipos de tratamentos foram restringidos em 2 principais: neo, que consiste em 1º quimioterapia e 2º cirurgia, ou adjuvante, que consiste em 1º cirurgia e 2º terapia. Gerando mais eficiência e possibilidade de revisão de diagnósticos.

Mais especificamente, o modelo deve utilizar técnicas de *machine learning*, testando sua acurácia e precisão para fornecer o melhor tratamento para pacientes diagnosticados com câncer de mama. Além disso, a ferramenta deve ser intuitiva e simples, para que os usuários (representados pelas *personas* na sessão 4.1.6.) possam usá-la com facilidade.

2.2. Proposta de Solução

A nossa proposta de solução envolve o consumo de dados que começaram a ser coletados a partir de 2008 de pacientes diagnosticados com câncer de mama. Através deles, será aplicado técnicas de *machine learning* para criação de modelos de classificações a fim de identificar o melhor tipo de tratamento (neo ou adjuvante), de acordo com o perfil e dados de cada paciente. Dessa forma, a classificação irá auxiliar os médicos responsáveis na decisão de qual tratamento recomendar ao paciente.

2.3. Justificativa

O uso de modelo preditivo é sem dúvidas uma excelente alternativa, pois o tratamento de câncer de mama se enquadra em casos que não sabemos exatamente o comportamento do fenômeno, ou seja, há uma grande influência da ótica de cada profissional de acordo com sua experiência. Com isso, como a IA trabalha diretamente com padrões, é possível ter uma acurácia pelo menos tão boa quanto a de profissionais formados. Além disso, a tecnologia apenas será utilizada para auxiliar na decisão, ou seja, a decisão final ainda será dos médicos, em que terão à disposição uma tecnologia que possibilitará ter mais assertividade na escolha do tratamento a sugerir.

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. Introdução

O câncer de mama é o tipo mais comum de câncer principalmente em mulheres ao redor do mundo. Ele se desenvolve quando as células da mama começam a crescer de forma descontrolada. Fatores de risco para o câncer de mama incluem idade avançada, histórico familiar de câncer de mama, exposição prolongada à radiação, obesidade, consumo excessivo de álcool e uso prolongado de terapia hormonal.

Esse tipo de câncer pode ser detectado cedo através de autoexame, mamografia e outros exames de imagem. Entretanto, esses métodos têm grande dependência de médicos e profissionais da saúde e estão suscetíveis a erros, atualmente cerca de 15% dos pacientes recebem um diagnóstico equivocado, além de existirem cerca de 3% de casos falsos positivos, que é quando o paciente é diagnosticado com câncer mas não é portador da doença.

O ICESP (Instituto de Câncer do Estado de São Paulo) é uma das unidades do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP), com atendimento exclusivo para pacientes da rede pública de saúde do SUS (Sistema Único de Saúde). Atualmente, o ICESP já atendeu mais de 121 mil pacientes e é referência nacional e internacional no tratamento contra o câncer

A utilização de inteligência artificial no diagnóstico de câncer tem sido cada vez mais frequente nos últimos anos, devido ao aumento da capacidade computacional e da quantidade de dados médicos disponíveis. Os impactos positivos de modelos preditivos incluem a possibilidade de detecção precoce do câncer, o que aumenta as chances de cura e tratamento eficaz. Além disso, esses modelos também podem ajudar os médicos a personalizar a abordagem de tratamento, identificando os pacientes mais propensos a desenvolver complicações ou a responder de forma inadequada a determinados tratamentos.4.1.1. Contexto da indústria

4.1.1.2. Forças de Porter

As forças de Porter são um framework de análise setorial que ajuda a entender o nível de competitividade de uma empresa inserida em um mercado específico. Usar esse modelo é vantajoso pois ele mapeia os fatores setoriais que impactam na empresa em questão. No caso da Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP) as 5 Forças de Porter são:

- Ameaça de novos concorrentes: O Hospital das Clínicas é referência em pesquisa na área da medicina, abrangendo as mais diversas áreas, mais especificamente na área de pesquisa contra o câncer de mama a instituição atualmente tem uma parceria com o ICESP (Instituto de Câncer do Estado de São Paulo) que tem infraestrutura de ponta. Entretanto, existem empresas que são possíveis concorrentes do FMUSP, principalmente dentre os escopos de universidades, podemos citar a Faculdade Albert Einstein, UFRGS (universidade federal do rio grande do sul) e a UNICAMP. Independente, a FMUSP ainda tem diversas vantagens que dependem de decretos do governo, como a liberação de verba, fato que faz a entrada de novos concorrentes mais complicada.

- Poder de negociação dos fornecedores: Atualmente o ICESP depende majoritariamente de verbas do governo, atualmente é o hospital que mais recebe verbas do governo, e por já contar com uma infraestrutura enorme, é extremamente improvável que o governo corte verbas do hospital, tanto que nos últimos anos a verba aumentou consideravelmente. No que tange aos fabricantes de equipamentos e fornecedores de maquinário, eles vêm das mais diversas fontes, normalmente importados do exterior. Ou seja, é possível afirmar que os fornecedores não têm poder de barganha suficiente para ameaçar o negócio.

- Poder de negociação com compradores/clientes: Os pacientes sempre irão buscar por hospitais de ponta para seu tratamento, portanto é possível afirmar que os pacientes seguirão tendo preferência pelo ICESP desde que o hospital siga sendo referência em pesquisa e tratamento de câncer no Brasil existirá um grande poder de negociação com compradores/clientes.

- Rivalidade entre concorrentes: Atualmente o ICESP concorre com outros hospitais contra o câncer, atualmente podemos destacar outros hospitais públicos que atendem pelo SUS como a Santa Casa da Misericórdia e HCPA, tanto quanto hospitais privados que prezam por excelência, como o Hospital Albert Einstein e Sírio Libanês.

- Ameaça de produtos substitutos: Atualmente existem diversos tratamentos sendo testados, a maior parte deles enfrenta um sério problema de custo, sendo absurdamente caros se comparados com os tratamentos mais convencionais. Dentre esses novos tratamentos podemos citar:
    - Imunoterapia: é um tratamento que incentiva o sistema imunológico a atacar exclusivamente as células cancerígenas, atualmente esse tratamento custa cerca de 50 mil reais por mês.
    - Terapia genética: o tratamento que introduz no organismo genes saudáveis - chamados de terapêuticos ou de interesse - para substituir, modificar ou

suplementar genes cancerosos, atualmente esse tratamento custa 2 milhões por aplicação.

4.1.1.3. Principais *players*

No Brasil existem diversos hospitais de referência no tratamento de câncer. Dessa forma, foi listado os 5 principais, além do próprio ICESP, sendo eles:

1. Hospital A.C. Camargo - Localizado em São Paulo, Brasil, o Hospital A.C. Camargo é considerado um dos melhores hospitais de tratamento de câncer do país. Com uma equipe altamente treinada de médicos e pesquisadores, o hospital oferece tratamentos personalizados para pacientes com câncer, incluindo cirurgias, radioterapia, quimioterapia e terapias-alvo.
2. Hospital Sírio-Libanês - Fundado em 1902, o Hospital Sírio-Libanês é uma instituição prestigiada em São Paulo, reconhecida por sua excelência em tratamentos de câncer. Com uma equipe altamente treinada e tecnologia avançada, o hospital oferece tratamentos personalizados para pacientes com câncer, incluindo cirurgias, radioterapia, quimioterapia e terapias-alvo.
3. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) - Fundado em 1980, o INCA é uma instituição nacional de pesquisa e tratamento de câncer no Brasil. Com uma equipe altamente treinada e colaboração estreita com instituições internacionais, o INCA tem como objetivo oferecer tratamentos inovadores e de alta qualidade para pacientes com câncer, além de contribuir para o avanço da pesquisa em oncologia.
4. Hospital Israelita Albert Einstein - Localizado em São Paulo, Brasil, o Hospital Israelita Albert Einstein é reconhecido como um dos melhores hospitais de tratamento de câncer do país. Com uma equipe multidisciplinar altamente treinada e tecnologia de ponta, o hospital oferece tratamentos personalizados para pacientes com câncer, incluindo cirurgias, radioterapia, quimioterapia, terapias-alvo e cuidados paliativos. Além disso, o hospital tem uma forte tradição em pesquisa e desenvolvimento de novos tratamentos e tecnologias na área de oncologia.
5. Hospital Santa Joana - Localizado em São Paulo, Brasil, o Hospital Santa Joana é reconhecido por sua excelência em tratamentos de câncer para mulheres. Com uma equipe de médicos especializados em oncologia ginecológica e obstétrica, o hospital oferece tratamentos personalizados e cuidados completos para pacientes com câncer de mama, ovários e outros tipos de câncer ginecológico. Além disso, o hospital tem uma forte tradição em pesquisa e educação na área de oncologia feminina.

4.1.1.4. Modelo de negócio

O ICESP - Instituto do Câncer do Estado de São Paulo é uma instituição pública de saúde do estado de São Paulo, com o objetivo de prestar atendimento oncológico gratuito à população. Portanto, seu modelo de negócio é baseado na prestação de serviços de saúde públicos, financiados pelo governo e pelos impostos pagos pela população.

4.1.1.5 .Tendências

Com o avanço da tecnologia e novos estudos, é possível notar algumas tendências no setor de tratamento de câncer de mama. Dessa forma, alguns exemplos de tendências são:

1. Terapias-alvo: cada vez mais, os tratamentos são direcionados aos tipos específicos de câncer, baseados nas características genéticas do tumor.
2. Imunoterapia: a imunoterapia tem se mostrado promissora no tratamento do câncer de mama, ajudando o sistema imunológico a combater as células cancerosas.
3. Cirurgia minimamente invasiva: a cirurgia minimamente invasiva, como a cirurgia de lumpectomia, está se tornando cada vez mais comum como opção de tratamento.
4. Terapia combinada: a combinação de vários tratamentos, como cirurgia, quimioterapia e radioterapia, é cada vez mais utilizada para obter resultados mais eficazes.
5. Utilização de tecnologia no auxílio da recomendação de qual o melhor tratamento para cada paciente, levando em consideração seus dados históricos médicos.

4.1.2. Análise SWOT

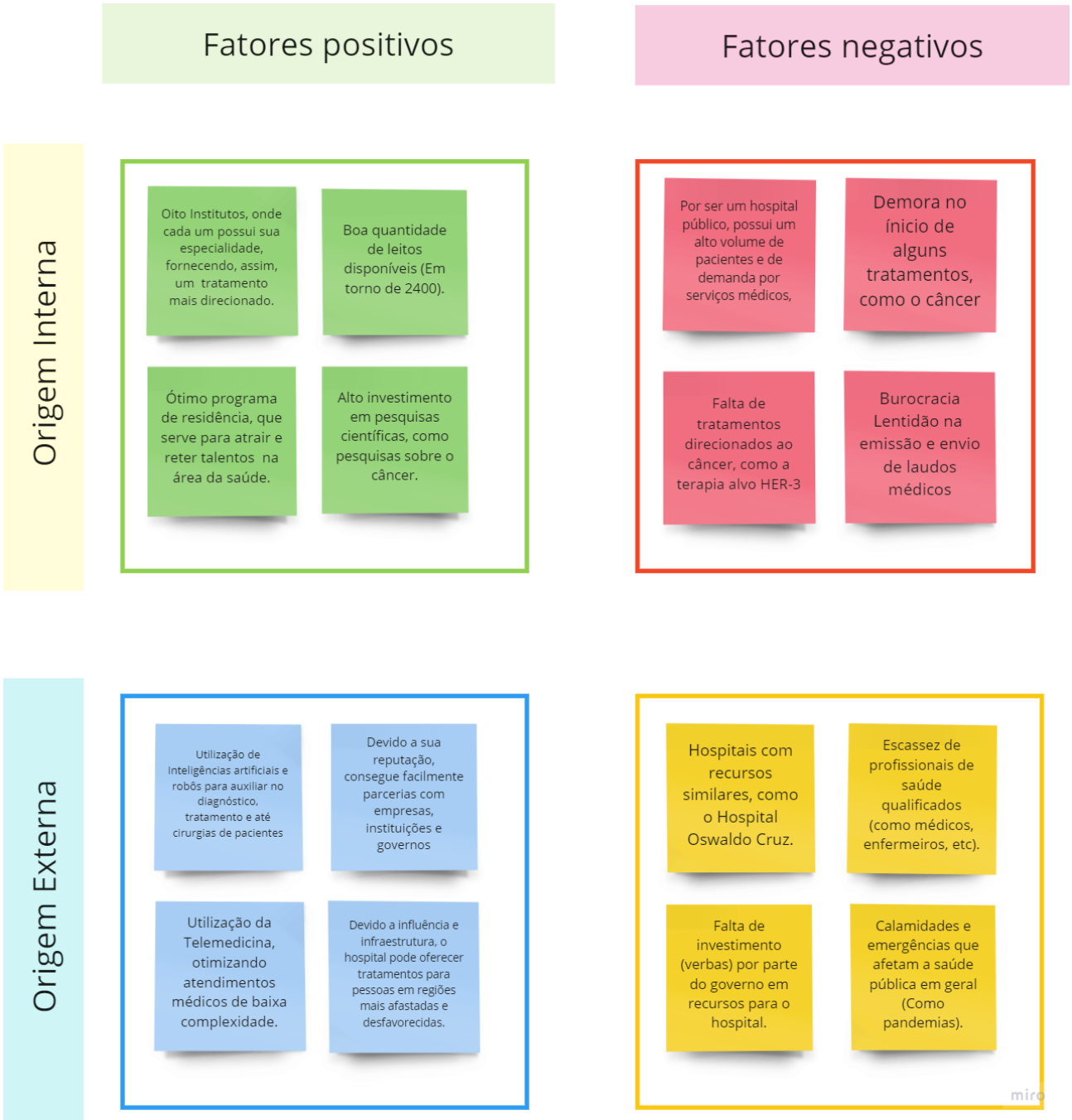
A análise SWOT é uma técnica de planejamento estratégico amplamente utilizada em empresas, organizações e projetos. A sigla SWOT significa Strengths (Forças), Weaknesses (Fraquezas), Opportunities (Oportunidades) e Threats (Ameaças). Nesse contexto, ela visa identificar as forças e fraquezas internas do projeto, bem como as oportunidades e ameaças externas que podem afetá-lo. É importante porque ajuda a compreender a situação atual do projeto e a identificar as melhores estratégias a seguir para atingir seus objetivos.

Em termos de benefícios, realizar uma análise SWOT pode trazer vários para o entendimento do contexto do negócio, incluindo:

- Identificação de pontos fortes e fracos: ao identificar as forças e fraquezas internas da empresa, é possível tomar medidas para maximizar as forças e corrigir as fraquezas.
- Identificação de oportunidades e ameaças: ao identificar as oportunidades e ameaças externas, é possível tomar medidas para aproveitar as oportunidades e minimizar as ameaças.
- Melhor tomada de decisões: a análise SWOT fornece informações importantes para ajudar na tomada de decisões estratégicas e de negócios.
- Melhor alinhamento de objetivos: a análise SWOT ajuda a garantir que todos os objetivos da empresa estejam alinhados com as forças, fraquezas, oportunidades e ameaças identificadas.
- Melhor entendimento do mercado: a análise SWOT permite entender melhor o mercado e as tendências que estão afetando ou poderão afetar a empresa no futuro.

Dessa forma, entendendo-se o que é, como é feito e sobre a importância de se fazer uma análise SWOT, foi elaborado uma em relação ao Hospital das Clínicas:

Figura 01:Análise SWOT do Hospital das Clínicas



Fonte: Elaboração dos autores.

Para melhor visualização, é possível acessar o link que irá redirecionar diretamente ao Miro: clique aqui.

4.1.3. Planejamento Geral da Solução

4.1.3.1. Qual é o problema a ser resolvido

A evolução do câncer de mama e sua resposta a tratamentos convencionais é muito variável. Dessa forma, o processo de decisão para definir qual o melhor tipo de tratamento para o paciente ainda possui muito da experiência pessoal dos médicos designados e *guidelines*, sendo necessário um suporte tecnológico para identificar padrões até então obscuros através dos dados clínicos fornecidos e informar o melhor tratamento para cada pessoa, com intuito de auxiliar os médicos na decisão de qual tratamento recomendar.

4.1.3.2. Qual a solução proposta (visão de negócios)

A solução à ser entregue será uma inteligência artificial que irá recomendar o melhor tratamento de acordo com o perfil de cada paciente, podendo ser o tratamento neo ou o adjuvante, atuando, assim, como um fator facilitador na decisão de um médico recomendar o melhor tratamento para o câncer de mama do paciente, ocasionando numa maior efetividade na recomendação, reduzindo possíveis gastos adicionais.

4.1.3.3. Como a solução proposta deverá ser utilizada

Os médicos ou enfermeiros terão acesso a uma plataforma web onde poderão fazer o *input* dos dados históricos e de exames do paciente já diagnosticado com câncer de mama, a fim de obter a melhor recomendação de tratamento e os motivos para determinada recomendação.

4.1.3.4. Quais os benefícios trazidos pela solução proposta

A solução proposta possui a vantagem de usar *Machine Learning* para processar uma quantidade enorme de dados que um ser humano não daria conta e, assim, estabelecer padrões, reduzindo a utilização de *guidelines* e experiências pessoais na recomendação de tratamento para os pacientes com câncer de mama. Dessa forma, a solução auxiliará na decisão final dos médicos.

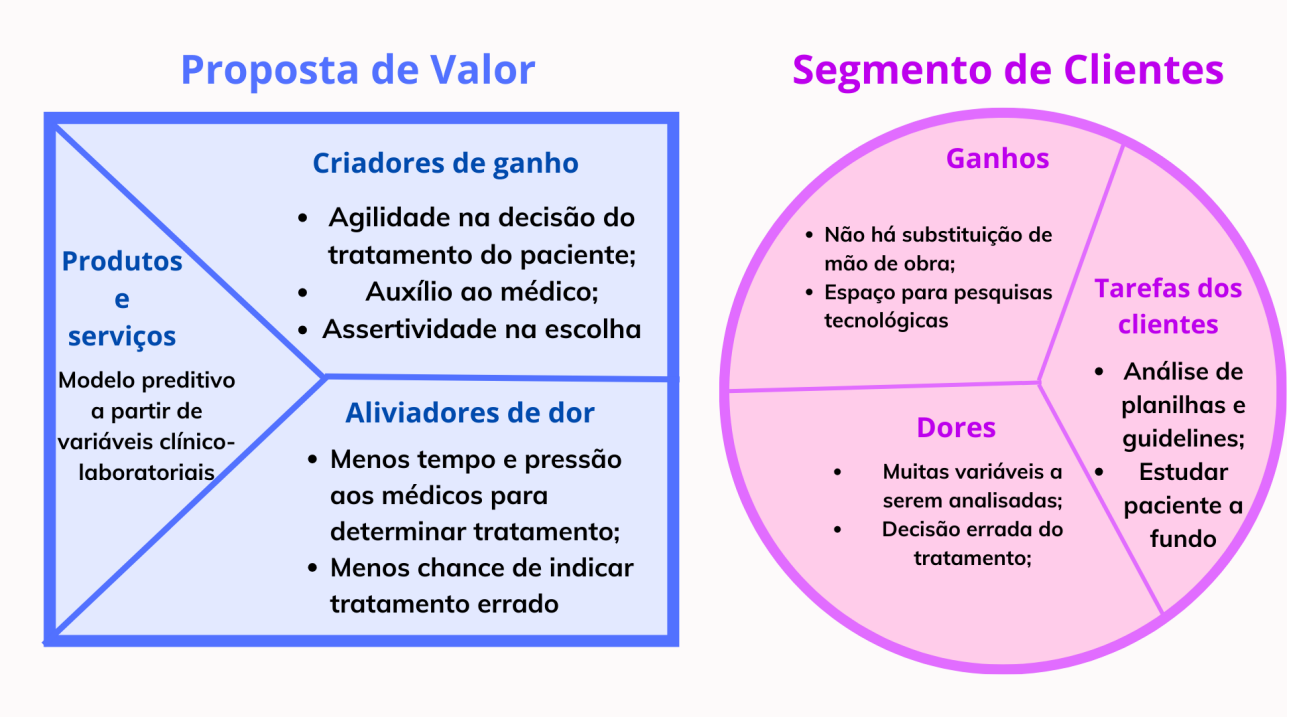
4.1.3.5. Qual será o critério de sucesso e qual medida será utilizada para o avaliar

O critério de sucesso do time será atingir um nível satisfatório de precisão e revocação, que consideramos ser maior que 80%. Assim, definimos o F1 score como a métrica mais adequada já que é a mais comumente usada em algoritmos de classificação binária, sendo precisamente a média harmônica entre a precisão (razão entre positivos verdadeiros e todos os positivos) e a revocação (divisão entre positivos verdadeiro pela soma deles com falsos negativos).

4.1.4. Value Proposition Canvas

O Value Proposition Canvas é uma ferramenta que ajuda a desenhar e entender a proposta de valor de um negócio. Nesse sentido, foi utilizado para visualizar e entender as necessidades, desejos e expectativas do parceiro e como podemos satisfazê-los de maneira única e valiosa.

Figura 02: Value Proposition Canvas do parceiro de negócios



Fonte: Elaboração dos autores.

Para melhor visualização, é possível acessar o link que irá redirecionar diretamente ao Canva: clique aqui.

4.1.5. Matriz de Riscos

A Matriz de Riscos é uma ferramenta de gerenciamento de riscos utilizada para identificar, avaliar e priorizar riscos de uma determinada iniciativa ou projeto. Nesse contexto, ela organiza os riscos em uma tabela que cruza a probabilidade de ocorrência de um risco com o seu impacto potencial. A matriz é então usada para priorizar os riscos, permitindo que a equipe tome decisões informadas sobre como lidar com eles.

Dessa forma, é possível, além de evitar problemas, criar oportunidades de preparação para algo que não pode ser evitado ou que possa impactar diretamente no resultado do projeto.

Figura 03: Matriz de Riscos do projeto

Probabilidade	Ameaças					Oportunidades					Possibilidade
90%			Enviesamento dos dados			Maior velocidade em escolha de tratamento					90%
70%						Possibilidade de uma "segunda opinião" em diagnósticos	Maior eficiência no trabalho de médicos		Menos estresse e carga de trabalho para médicos		70%
50%				Dificuldade na interpretração do modelo	Escolha errada no tratamento	Maior precisão na escolha do tratamento					50%
30%				Falta de química no grupo							30%
10%	Pouca velocidade para execução do modelo				Dados insuficientes para realizar um modelo preciso						10%
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo	

Fonte: Elaboração dos autores.

Para melhor visualização, é possível acessar o link que irá redirecionar diretamente ao Sheets: clique aqui.

4.1.6. Personas

Persona é um personagem fictício que representa o cliente potencial de um negócio ou projeto. É baseado em dados e características de clientes reais, como comportamento, dados demográficos, problemas, desafios e objetivos. A persona é uma ferramenta de segmentação de mercado, usada para guiar a tomada de decisões de design, desenvolvimento de produtos e marketing. Com isso, garante que a equipe esteja sempre alinhada aos interesses e necessidades dos usuários, conseguindo identificar oportunidades para melhorar a experiência e aumentar a satisfação deles.

4.1.6.1. Persona que utiliza o modelo:

Figura 04: Persona 'Médica'

Dra Carolina Santos

CARACTERÍSTICAS

- Experiente e dedicada, tem quase 15 anos de experiência na área de oncologia.
- Ela acredita que todos têm o direito de viver e curar o câncer com dignidade.
- No seu tempo livre, dedica-se à sua família.

Idade: 45

Profissão: Médica Oncologista

Localidade: São Paulo - SP

FRUSTRAÇÕES

- Quando a resposta dos tratamentos não é satisfatória.
- Quando demora para indicar o tratamento.
- Falta de estrutura em muitos hospitais;
- Polarização política no país.
- Frustrada com o aumento de quantidade de informações falsas sobre medicina nas redes sociais.

METAS

- Melhorar a precisão e eficácia dos tratamentos de câncer oferecidos.
- Escrever uma autobiografia.
- Virar referência na área de oncologia no Brasil.
- Aumentar a sobrevida de pacientes com câncer de mama.
- Viajar por toda a Europa.

Fonte: Elaboração dos autores.

4.1.6.2. Persona afetada pela solução:

Figura 05: Persona 'Paciente'

Fernanda Silveira

CARACTERÍSTICAS

- Fernanda é uma pessoa extremamente gentil que adora lidar com crianças e adolescentes
- Possui 2 filhos nos quais é apaixonada, chamados Vitória e Rafael.
- Gosta de passar o tempo lendo e trabalhando.
- Foi diagnosticada com câncer de mama recentemente, deixando totalmente apreensiva por não confiar que o tratamento fará efeito

Idade: 48

Profissão: Professora

Localidade: São Paulo - SP

FRUSTRAÇÕES

- Frustrada com o diagnóstico de câncer;
- Ansiosa com seu futuro incerto
- Medo de não ver seus filhos se formarem;
- Medo do tratamento não fazer efeito;
- Medo de ter que retirar as mamas;

METAS

- Se curar do câncer de mama;
- Ser avó;
- Reformar a casa no interior para poder morar;
- Liberdade financeira;
- Viajar para outro país.

Fonte: Elaboração dos autores.

Para melhor visualização das *personas*, é possível acessar o link que irá redirecionar diretamente ao Canva: clique aqui.

4.1.7. Jornadas do Usuário

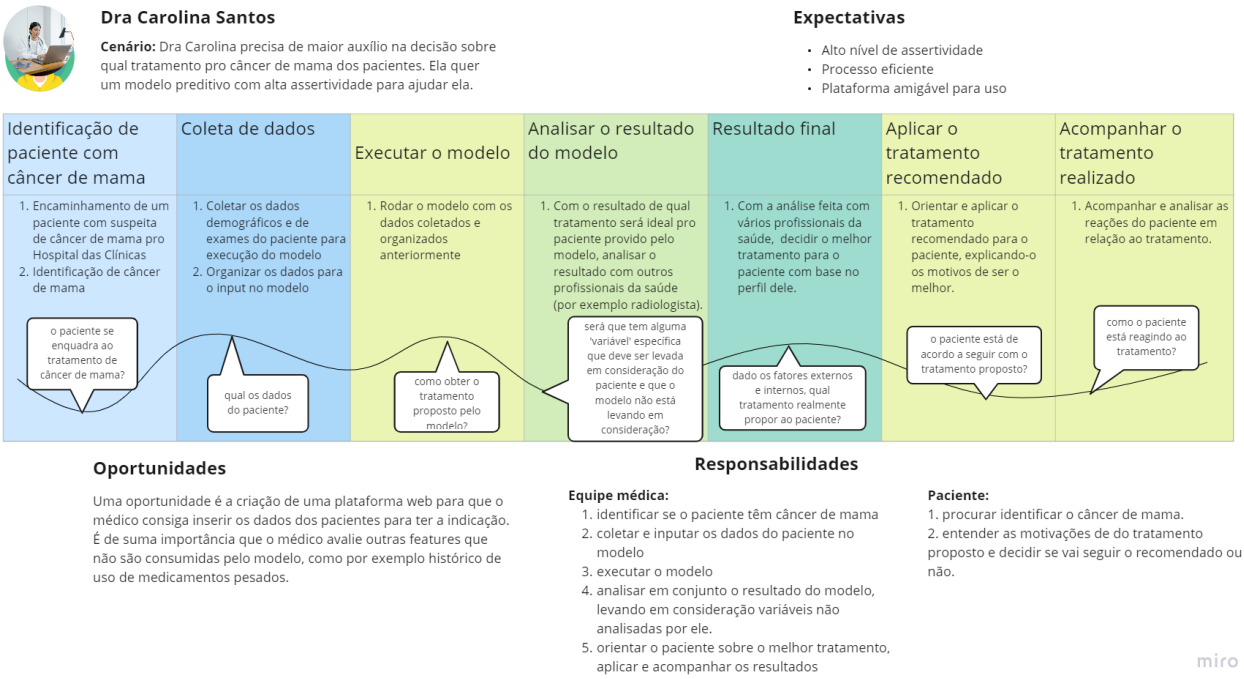
Um mapa de jornada é uma visualização do processo pelo qual uma pessoa passa para atingir um objetivo. O mapeamento da jornada começa compilando uma série de ações do usuário em uma linha do tempo. Em seguida, a linha do tempo é desenvolvida com pensamentos e emoções do usuário para criar uma narrativa.

Com isso, foi utilizado porque permite entender as dores, desafios, motivações e expectativas dos clientes ao longo da jornada, fazendo com que seja possível aprimorar a experiência do usuário e, assim, aumentar a satisfação das *personas*.

Para melhor visualização das Jornadas do Usuário, é possível acessar o link que irá redirecionar diretamente ao Miro: clique aqui

4.1.7.1. Jornadas do Usuário da Médica

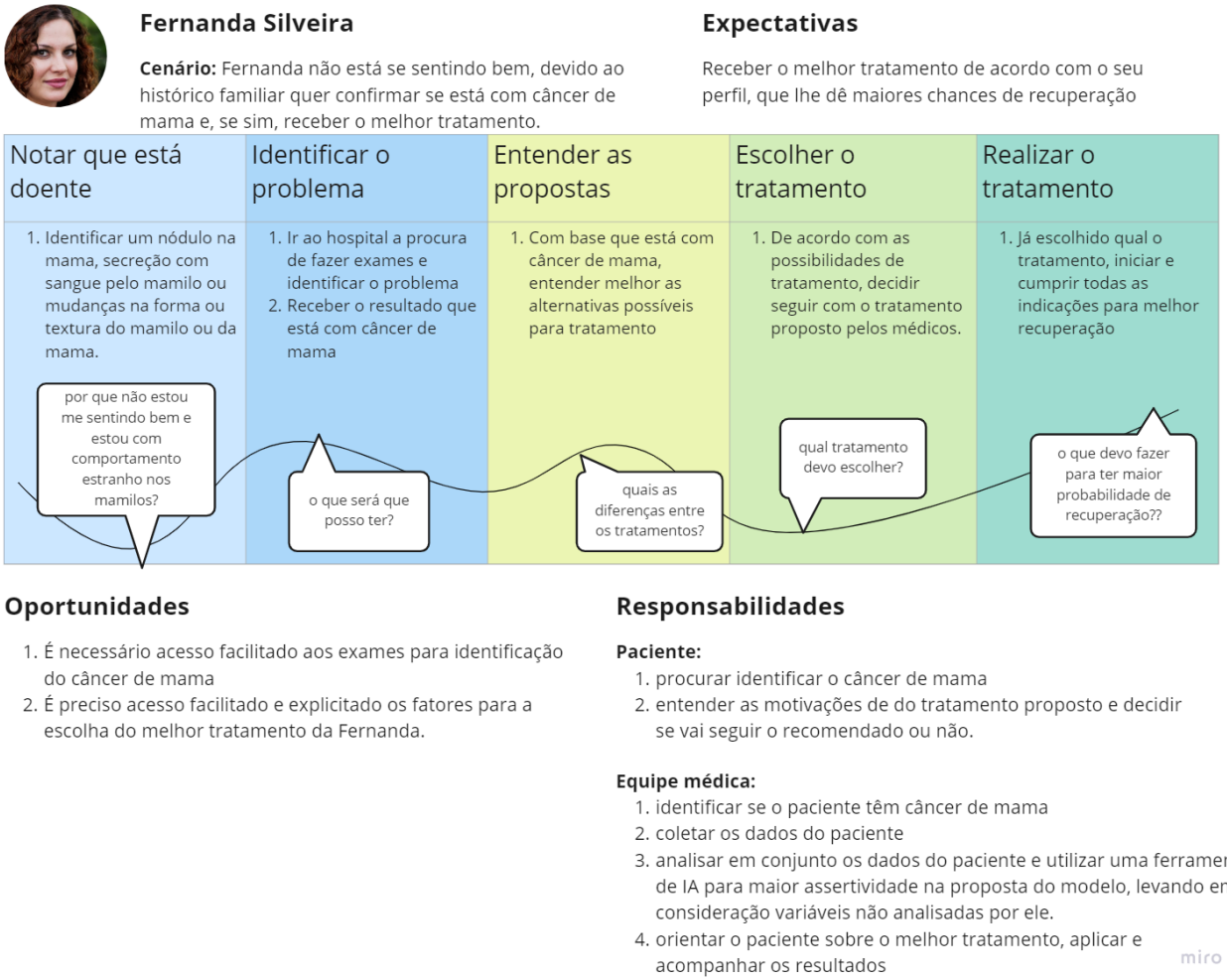
Figura 06: Jornada do Usuário da Médica



Fonte: Elaboração dos autores.

4.1.7.2. Jornadas do Usuário da Paciente

Figura 07: Jornada do Usuário da Paciente



Fonte: Elaboração dos autores.

4.1.8. User Stories

User Story ou “história de usuário” é uma descrição concisa de uma necessidade do usuário do produto (ou seja, de um “requisito”) sob o ponto de vista deste usuário. A User Story busca descrever essa necessidade de uma forma simples e leve, garantindo que a equipe esteja alinhada aos interesses reais dos usuários.

As *User Stories* foram organizadas e ordenadas de acordo com a priorização, esforço e risco. A escala utilizada para elencar os atributos de esforço, risco e impacto varia de 1 a 5, sendo 1 considerado ‘nenhum’ e 5 considerado ‘muito’, enquanto para priorização, foi dividido entre alta, média e baixa.

Tabela 01: História do usuário da paciente

HISTÓRIA DO USUÁRIO	ESFORÇO	RISCO	IMPACTO
PRIORIZAÇÃO ALTA			
Como Fernanda, quero que o hospital utilize um modelo preditivo que ajude a escolher o melhor tratamento para o meu câncer de mama, para ter maior confiança na minha recuperação.	5	4	5
PRIORIZAÇÃO MÉDIA			
Como Fernanda, quero ter acesso a informações sobre os efeitos colaterais dos tratamentos recomendados para mim, para que eu possa ter uma ideia de quais são as melhores opções para mim.	3	2	3
PRIORIZAÇÃO BAIXA			
Como Fernanda, quero ter acesso aos motivos de determinado tratamento ser recomendado para mim, para que eu possa ter mais confiança na escolha dos tratamentos.	3	2	2

Fonte: Elaboração dos autores.

Tabela 02: História do usuário da médica

HISTÓRIA DO USUÁRIO	ESFORÇO	RISCO	IMPACTO
PRIORIZAÇÃO ALTA			
Como Médica, quero ter acesso a um modelo preditivo que me auxilie sobre qual conjunto de tratamentos será melhor para o paciente para que minha decisão final seja mais embasada.	5	4	5
Como Médica, quero entender como o modelo preditivo chega às suas recomendações, para que eu possa ter certeza de que ele está levando em conta todos os fatores relevantes.	4	4	4

PRIORIZAÇÃO MÉDIA			
Como Médica, quero ter acesso a históricos de dados tratados de pacientes, para conseguir visualizar caso a caso e entender o comportamento da recuperação de determinado paciente.	3	2	3
Como Médica, quero ter acesso a uma plataforma web para ser capaz de imputar dados manualmente ou massivamente, para que o modelo consiga me ajudar no dia a dia.	4	4	3
PRIORIZAÇÃO BAIXA			
Como Médica, quero ser capaz de compartilhar as informações obtidas pelo modelo preditivo com o meu paciente, para que eu possa explicá-lo melhor sobre o processo de decisão.	3	2	2

Fonte: Elaboração dos autores.

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

NeoVision, pessoa jurídica de direito privado leva a sua privacidade a sério e zela pela segurança e proteção de dados de todos os seus clientes, parceiros, fornecedores e usuários do site domínio https://www.neovision.com.br/politica-de-privacidade e qualquer outro site, ferramenta ou aplicativo operado pela empresa.

Esta Política de Privacidade destina-se a informá-lo sobre o modo como nós utilizamos e divulgamos informações coletadas em suas visitas ao nosso modelo preditivo.

Esta Política de Privacidade aplica-se somente a informações coletadas por meio da empresa.

AO ACESSAR O MODELO PREDITIVO, ENVIAR COMUNICAÇÕES OU FORNECER QUALQUER TIPO DE DADO PESSOAL, VOCÊ DECLARA ESTAR CIENTE COM RELAÇÃO AOS TERMOS AQUI PREVISTOS E DE ACORDO COM A POLÍTICA DE PRIVACIDADE, A QUAL DESCREVE AS FINALIDADES E FORMAS DE TRATAMENTO DE SEUS DADOS PESSOAIS QUE VOCÊ DISPONIBILIZAR NA COMPANHIA.

Esta Política de Privacidade fornece uma visão geral de nossas práticas de privacidade e das escolhas que você pode fazer, bem como direitos que você pode exercer em relação aos Dados Pessoais tratados por nós. Se você tiver alguma dúvida sobre o uso de Dados Pessoais, entre em contato com neovision.mod03@gmail.com.

Além disso, a Política de Privacidade não se aplica a quaisquer aplicativos, produtos, serviços, site ou recursos de mídia social de terceiros que possam ser oferecidos ou acessados por meio da companhia. O acesso a esses links fará com que você deixe o nosso site e poderá resultar na coleta ou compartilhamento de informações sobre você por terceiros. Nós não controlamos, endossamos ou fazemos quaisquer representações sobre sites de terceiros ou suas práticas de privacidade, que podem ser diferentes das nossas. Recomendamos que você revise a política de privacidade de qualquer site com o qual você interaja antes de permitir a coleta e o uso de seus Dados Pessoais.

Caso você nos envie Dados Pessoais referentes a outras pessoas físicas, você declara ter a competência para fazê-lo e declara ter obtido o consentimento necessário para autorizar o uso de tais informações nos termos desta Política de Privacidade.

Seção 1 - Definições

Para os fins desta Política de Privacidade:

1. "Dados Pessoais": significa qualquer informação que, direta ou indiretamente, identifique ou possa identificar uma pessoa natural, como por exemplo, nome, CPF, data de nascimento, endereço IP, dentre outros;
2. "Dados Pessoais Sensíveis": significa qualquer informação que revele, em relação a uma pessoa natural, origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico;
3. "Tratamento de Dados Pessoais": significa qualquer operação efetuada no âmbito dos Dados Pessoais, por meio de meios automáticos ou não, tal como a recolha, gravação, organização, estruturação, armazenamento, adaptação ou alteração, recuperação, consulta, utilização, divulgação por transmissão, disseminação ou, alternativamente, disponibilização, harmonização ou associação, restrição, eliminação ou destruição. Também é considerado Tratamento de Dados Pessoais qualquer outra operação prevista nos termos da legislação aplicável;

4. "Leis de Proteção de Dados": significa todas as disposições legais que regulam o Tratamento de Dados Pessoais, incluindo, porém sem se limitar, a Lei nº 13.709/18, Lei Geral de Proteção de Dados Pessoais ("LGPD").

Seção 2 - Uso de Dados Pessoais

Coletamos e usamos Dados Pessoais para gerenciar seu relacionamento conosco e melhor atendê-lo quando você estiver adquirindo produtos e/ou serviços na companhia, personalizando e melhorando sua experiência. Exemplos de como usamos os dados incluem:

1. Viabilizar que você adquira produtos e/ou serviços na companhia;
2. Para confirmar ou corrigir as informações que temos sobre você;
3. Para enviar informações que acreditamos ser do seu interesse;
4. Para personalizar sua experiência de uso da companhia;
5. Para entrarmos em contato por um número de telefone e/ou endereço de e-mail fornecido. Podemos entrar em contato com você pessoalmente, por mensagem de voz, através de equipamentos de discagem automática, por mensagens de texto (SMS), por e-mail, ou por qualquer outro meio de comunicação que seu dispositivo seja capaz de receber, nos termos da lei e para fins comerciais razoáveis.

Além disso, os Dados Pessoais fornecidos também podem ser utilizados na forma que julgarmos necessária ou adequada: (a) nos termos das Leis de Proteção de Dados; (b) para atender exigências de processo judicial; (c) para cumprir decisão judicial, decisão regulatória ou decisão de autoridades competentes, incluindo autoridades fora do país de residência; (d) para aplicar nossos Termos e Condições de Uso; (e) para proteger nossas operações; (f) para proteger direitos, privacidade, segurança nossos, seus ou de terceiros; (g) para detectar e prevenir fraude; (h) permitir-nos usar as ações disponíveis ou limitar danos que venhamos a sofrer; e (i) de outros modos permitidos por lei.

Seção 3 - Não fornecimento de Dados Pessoais

Não há obrigatoriedade em compartilhar os Dados Pessoais que solicitamos. No entanto, se você optar por não os compartilhar, em alguns casos, não poderemos fornecer a você acesso completo à companhia, alguns recursos especializados ou ser capaz de prestar a assistência necessária ou, ainda, viabilizar a entrega do produto ou prestar o serviço contratado por você.

Seção 4 - Dados coletados

O público em geral poderá navegar na companhia sem necessidade de qualquer cadastro e envio de Dados Pessoais. No entanto, algumas das funcionalidades da companhia poderão depender de cadastro e envio de Dados Pessoais como concluir a compra/contratação do serviço e/ou a viabilizar a entrega do produto/prestação do serviço por nós.

No contato a companhia, nós podemos coletar:

1. Dados de contato: nome, sobrenome, número de telefone, endereço, cidade, estado e endereço de e-mail;
2. Informações enviadas: informações que você envia via formulário (dúvidas, reclamações, sugestões, críticas, elogios etc.).

Na navegação geral na companhia, nós poderemos coletar:

1. Dados de localização: dados de geolocalização quando você acessa a companhia;
2. Preferências: informações sobre suas preferências e interesses em relação aos produtos/serviços (quando você nos diz o que eles são ou quando os deduzimos do que sabemos sobre você);
3. Dados de navegação na companhia: informações sobre suas visitas e atividades, incluindo o conteúdo (e quaisquer anúncios) com os quais você visualiza e interage, informações sobre o navegador e o dispositivo que você está usando, seu endereço IP, sua localização, o endereço do site a partir do qual você chegou. Algumas dessas informações são coletadas usando nossas Ferramentas de Coleta Automática de Dados, que incluem cookies, web beacons e links da web incorporados. Para saber mais, leia como nós usamos Ferramentas de Coleta Automática de Dados na seção 7 abaixo;
4. Dados anônimos ou agregados: respostas anônimas para pesquisas ou informações anônimas e agregadas sobre como a companhia é usufruída. Durante nossas operações, em certos casos, aplicamos um processo de desidentificação ou pseudonimização aos seus dados para que seja razoavelmente improvável que você identifique você através do uso desses dados com a tecnologia disponível;
5. Outras informações que podemos coletar: informações que não revelem especificamente a sua identidade ou que não são diretamente relacionadas a um indivíduo, tais como informações sobre navegador e dispositivo; dados de uso da companhia; e informações coletadas por meio de cookies, pixel tags e outras tecnologias.
6. Dados pessoais sensíveis: dados que permitem identificar de forma direta ou indireta uma determinada pessoa, tais como nome, CPF, RG, carteira de habilitação, passaporte,

número de telefone, endereço, e-mail, IP e até cookies. Além disso, dados históricos clínicos de pacientes.

Nós não coletamos Dados Pessoais Sensíveis durante a navegação geral.

Seção 5 - Compartilhamento de Dados Pessoais com terceiros

Nós poderemos compartilhar seus Dados Pessoais:

1. Com a(s) empresa(s) parceira(s) que você selecionar ou optar em enviar os seus dados, dúvidas, perguntas etc., bem como com provedores de serviços ou parceiros para gerenciar ou suportar certos aspectos de nossas operações comerciais em nosso nome. Esses provedores de serviços ou parceiros podem estar localizados nos Estados Unidos, na Argentina, no Brasil ou em outros locais globais, incluindo servidores para homologação e produção, e prestadores de serviços de hospedagem e armazenamento de dados, gerenciamento de fraudes, suporte ao cliente, vendas em nosso nome, atendimento de pedidos, personalização de conteúdo, atividades de publicidade e marketing (incluindo publicidade digital e personalizada) e serviços de TI, por exemplo;
2. Com terceiros, com o objetivo de nos ajudar a gerenciar a companhia;
3. Com terceiros, caso ocorra qualquer reorganização, fusão, venda, joint venture, cessão, transmissão ou transferência de toda ou parte da nossa empresa, ativo ou capital (incluindo os relativos à falência ou processos semelhantes).

Seção 6 - Transferências internacionais de dados

Dados Pessoais e informações de outras naturezas coletadas por nós podem ser transferidos ou acessados por entidades pertencentes ao grupo corporativo das empresas parceiras em todo o mundo de acordo com esta Política de Privacidade.

Seção 7 - Coleta automática de Dados Pessoais

Quando você visita a companhia, ela pode armazenar ou recuperar informações em seu navegador, principalmente na forma de cookies, que são arquivos de texto contendo pequenas quantidades de informação. Essas informações podem ser sobre você, suas preferências ou seu dispositivo e são usadas principalmente para que a companhia funcione como você espera. As informações geralmente não o identificam diretamente, mas podem oferecer uma experiência na internet mais personalizada.

De acordo com esta Política de Privacidade, nós e nossos prestadores de serviços terceirizados, mediante seu consentimento, podemos coletar seus Dados Pessoais de diversas formas, incluindo, entre outros:

1. Por meio do navegador ou do dispositivo: algumas informações são coletadas pela maior parte dos navegadores ou automaticamente por meio de dispositivos de acesso à internet, como o tipo de computador, resolução da tela, nome e versão do sistema operacional, modelo e fabricante do dispositivo, idioma, tipo e versão do navegador de Internet que está utilizando. Podemos utilizar essas informações para assegurar que a companhia funcione adequadamente.
2. Uso de cookies: informações sobre o seu uso da companhia podem ser coletadas por terceiros a partir de cookies. Cookies são informações armazenadas diretamente no computador que você está utilizando. Os cookies permitem a coleta de informações tais como o tipo de navegador, o tempo despendido na companhia, as páginas visitadas, as preferências de idioma, e outros dados de tráfego anônimos. Nós e nossos prestadores de serviços utilizamos informações para proteção de segurança, para facilitar a navegação, exibir informações de modo mais eficiente, e personalizar sua experiência ao utilizar a companhia, assim como para rastreamento online. Também coletamos informações estatísticas sobre o uso da companhia para aprimoramento contínuo do nosso design e funcionalidade, para entender como a companhia é utilizada e para auxiliá-lo a solucionar questões relativas à companhia.

   Caso não deseje que suas informações sejam coletadas por meio de cookies, há um procedimento simples na maior parte dos navegadores que permite que os cookies sejam automaticamente rejeitados, ou oferece a opção de aceitar ou rejeitar a transferência de um cookie (ou cookies) específico(s) de um site determinado para o seu computador. Entretanto, isso pode causar problemas no uso do software da empresa.

   As definições que você escolher podem afetar sua experiência de navegação e o funcionamento que depende da utilização de cookies. Nesse sentido, não nos responsabilizamos pelas consequências decorrentes do funcionamento limitado da empresa causado pela desativação de cookies em seu dispositivo (incapacidade de definir ou ler um cookie).
3. Uso de pixel tags e outras tecnologias similares: pixel tags (também conhecidos como Web beacons e GIFs invisíveis) podem ser utilizados para rastrear ações de usuários da

companhia (incluindo destinatários de e-mails), medir o sucesso das nossas campanhas de marketing e coletar dados estatísticos sobre o uso da companhia e taxas de resposta, e ainda para outros fins não especificados. Podemos contratar empresas de publicidade comportamental, para obter relatórios sobre os anúncios da companhia em toda a internet. Para isso, essas empresas utilizam cookies, pixel tags e outras tecnologias para coletar informações sobre a sua utilização, ou sobre a utilização de outros usuários, da nossa companhia e de sites de terceiros. Nós não somos responsáveis por pixel tags, cookies e outras tecnologias similares utilizadas por terceiros.

Seção 8 - Categorias de cookies

Os cookies utilizados na nossa companhia estão de acordo com os requisitos legais e são enquadrados nas seguintes categorias:

1. Estritamente necessários: estes cookies permitem que você navegue pelo site e desfrute de recursos essenciais com segurança. Um exemplo são os cookies de segurança, que autenticam os usuários, protegem os seus dados e evitam a criação de logins fraudulentos.
2. Desempenho: os cookies desta categoria coletam informações de forma codificada e anônima relacionadas à nossa companhia virtual, como, por exemplo, o número de visitantes de uma página específica, origem das visitas ao site e quais as páginas acessadas pelo usuário. Todos os dados coletados são utilizados apenas para eventuais melhorias no site e para medir a eficácia da nossa comunicação.
3. Funcionalidade: estes cookies são utilizados para lembrar definições de preferências do usuário com o objetivo de melhorar a sua visita no nosso site, como, por exemplo, configurações aplicadas no layout do site ou suas respostas para pop-ups de promoções e cadastros -; dessa forma, não será necessário perguntar inúmeras vezes.
4. Publicidade: utilizamos cookies com o objetivo de criar campanhas segmentadas e entregar anúncios de acordo com o seu perfil de consumo na nossa companhia virtual.

Seção 9 - Direitos do Usuário

Você pode, a qualquer momento, requerer: (i) confirmação de que seus Dados Pessoais estão sendo tratados; (ii) acesso aos seus Dados Pessoais; (iii) correções a dados incompletos, inexatos ou desatualizados; (iv) anonimização, bloqueio ou eliminação de dados desnecessários, excessivos ou tratados em desconformidade com o disposto em lei; (v) portabilidade de Dados Pessoais a outro prestador de serviços, contanto que isso não afete nossos segredos industriais e tecnológicos; (vi) eliminação de Dados Pessoais tratados com seu consentimento, na medida do permitido em lei; (vii) informações sobre as entidades às quais seus Dados Pessoais tenham sido

compartilhados; (viii) informações sobre a possibilidade de não fornecer o consentimento e sobre as consequências da negativa; e (ix) revogação do consentimento. Os seus pedidos serão tratados com especial cuidado de forma a que possamos assegurar a eficácia dos seus direitos. Poderá lhe ser pedido que faça prova da sua identidade de modo a assegurar que a partilha dos Dados Pessoais é apenas feita com o seu titular.

Você deverá ter em mente que, em certos casos (por exemplo, devido a requisitos legais), o seu pedido poderá não ser imediatamente satisfeito, além de que nós poderemos não conseguir atendê-lo por conta de cumprimento de obrigações legais.

Seção 10 - Segurança dos Dados Pessoais

Buscamos adotar as medidas técnicas e organizacionais previstas pelas Leis de Proteção de Dados adequadas para proteção dos Dados Pessoais na nossa organização. Infelizmente, nenhuma transmissão ou sistema de armazenamento de dados tem a garantia de serem 100% seguros. Caso tenha motivos para acreditar que sua interação conosco tenha deixado de ser segura (por exemplo, caso acredite que a segurança de qualquer uma de seus dados foi comprometida), favor nos notificar imediatamente.

Seção 11 - Links de hipertexto para outros sites e redes sociais

A companhia poderá, de tempos a tempos, conter links de hipertexto que redirecionará você para sites das redes dos nossos parceiros. Se você clicar em um desses links para qualquer um desses sites, lembre-se que cada site possui as suas próprias práticas de privacidade e que não somos responsáveis por essas políticas. Consulte as referências políticas antes de enviar quaisquer Dados Pessoais para esses sites.

Não nos responsabilizamos pelas políticas e práticas de coleta, uso e divulgação (incluindo práticas de proteção de dados) de outras organizações, tais como Facebook, Apple, Google, Microsoft, ou de qualquer outro desenvolvedor de software ou provedor de aplicativo, companhia de mídia social, sistema operacional, prestador de serviços de internet sem fio ou fabricante de dispositivos, incluindo todos os Dados Pessoais que divulgar para outras organizações por meio dos aplicativos, relacionadas a tais aplicativos, ou publicadas em nossas páginas em mídias sociais. Nós recomendamos que você se informe sobre a política de privacidade de cada site visitado ou de cada prestador de serviço utilizado.

Seção 12 - Atualizações desta Política de Privacidade

Se modificarmos nossa Política de Privacidade, publicaremos o novo texto na companhia, com a data de revisão atualizada. Podemos alterar esta Política de Privacidade a qualquer momento. Caso haja alteração significativa nos termos desta Política de Privacidade, podemos informá-lo por meio das informações de contato que tivermos em nosso banco de dados ou por meio de notificação em nossa companhia.

Recordamos que nós temos como compromisso não tratar os seus Dados Pessoais de forma incompatível com os objetivos descritos acima, exceto se de outra forma requerido por lei ou ordem judicial.

Sua utilização da companhia após as alterações significa que aceitou as Políticas de Privacidade revisadas. Caso, após a leitura da versão revisada, você não esteja de acordo com seus termos, favor encerrar o acesso à companhia.

Seção 13 - Encarregado do tratamento dos Dados Pessoais

Caso pretenda exercer qualquer um dos direitos previstos, inclusive retirar o seu consentimento, nesta Política de Privacidade e/ou nas Leis de Proteção de Dados, ou resolver quaisquer dúvidas relacionadas ao Tratamento de seus Dados Pessoais, favor contatar-nos em neovision.mod03@gmail.com.

4.2. Compreensão dos Dados

1. Exploração de dados:

A exploração de dados é a etapa em que se investiga os dados para entender melhor sua estrutura e padrões, identificar problemas e tendências e encontrar relações entre as variáveis. Isso ajuda a construir um modelo mais preciso e confiável, identificando possíveis desafios e oportunidades nos dados antes de criar o modelo.

Para exploração dos dados, foi utilizado o 'Profile Report' que gera um relatório que inclui estatísticas descritivas para cada variável, como número de valores ausentes, valores únicos, distribuição, entre outras informações relevantes para compreender melhor os dados. Nesse contexto, para auxiliar no controle da exploração, foi realizado o preenchimento da atuação, explicação da coluna e sua importância na seguinte tabela Sheets: clique aqui para acessar a tabela.

Nesse sentido, para acessar os relatórios é possível através dos htmls:

NOME DA TABELA	LINK PARA ACESSAR O RELATÓRIO
HISTOPATOLOGIA	profile_histo.html
DEMOGRÁFICO	profile_demo.html
REGISTRO DE TUMOR	profile_reg_tumo.html
PESO E ALTURA	profile_peso_alt.html

a) Cite quais são as colunas numéricas e categóricas.

TABELA HISTOPATOLOGIA	
nome da coluna	**tipo da coluna**
Record ID	numérica
Repeat Instrument	numérica
Repeat Instance	numérica
Diagnostico primario (tipo histológico)	categórico
Grau histológico	categórico
Subtipo tumoral	categórico
Receptor de estrogênio	categórico
Receptor de progesterona	categórico
Ki67 (>14%)	categórico

Receptor de progesterona (quantificação %)	categórico
Receptorde Estrogênio (quantificação %)	categórico
Indice H (Receptorde progesterona)	numérica
HER2 por IHC	categórico
HER2 por FISH	categórico
Ki67 (%)	numérica

TABELA DEMOGRÁFICO	
nome da coluna	**tipo da coluna**
Record ID	numérica
Repeat Instrument	categórico
Repeat Instance	numérica
Escolaridade	categórico
Idade do paciente ao primeiro diagnóstico	numérica
Sexo	categórico
Raça declarada (Biobanco)	categórico
UF de nascimento do paciente	categórico
UF de residência do paciente	categórico
Data da última informação sobre o paciente	datetime
Última informação do paciente	categórico
Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci]	numérica
Já ficou grávida?	categórico
Quantas vezes ficou gráfida?	quantitativo
Número de partos	quantitativo
Idade na primeira gestação	quantitativo
Abortou	categórico
Amamentou na primeira gestação?	categórico
Por quanto tempo amamentou?	quantitativo
Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Não)	categórico
Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 1º grau, apenas 1 caso)	categórico
Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 1º grau, mais de 1	categórico

caso)	
Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, apenas 1 caso)	categórico
Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, mais de 1 caso)	categórico
Idade da primeira mentruação	quantitativo
Faz uso de métodos contraceptivo?	categórico
Qual método? (choice=Pílula anticoncepcional)	categórico
Qual método? (choice=DIU)	categórico
Qual método? (choice=camisinha)	categórico
Qual método? (choice=outros)	categórico
Qual método? (choice=não informou)	categórico
Já fez uso de drogas?	categórico
Atividade Física	categórico
Consumo de tabaco	categórico
Consumo de álcool	categórico
Possui histórico familiar de câncer?	categórico
Grau de parentesco de familiar com cancer? (choice=primeiro (pais, irmãos, filhos))	categórico
Grau de parentesco de familiar com cancer? (choice=segundo (avós, tios e netos))	categórico
Grau de parentesco de familiar com cancer? (choice=terceiro (bisavós, tio avós, primos, sobrinhos))	categórico
Regime de Tratamento	categórico
Hormonioterapia	categórico
Data da cirurgia	datetime
Tipo de terapia anti-HER2 neoadjuvante	categórico
Radioterapia	categórico
Data de início do tratamento quimioterapia	data
Esquema de hormonioterapia	categórico
Data do início Hormonioterapia adjuvante	datetime
Data de início da Radioterapia	datetime
TABELA REGISTRO DE TUMORES	
nome da coluna	**tipo da coluna**
Record ID	numérica
Repeat Instrument	categórico

Repeat Instance	numérica
Data da primeira consulta institucional [dt_pci]	datetime
Data do diagnóstico	datetime
Código da Topografia (CID-O)	categórico
Código da Morfologia de acordo com o CID-O	categórico
Estadio Clínico	categórico
Grupo de Estadio Clínico	categórico
Classificação TNM Clínico - T	categórico
Classificação TNM Clínico - N	categórico
Classificação TNM Clínico - M	categórico
Metastase ao DIAGNOSTICO - CID-O #1	categórico
Metastase ao DIAGNOSTICO - CID-O #2	categórico
Metastase ao DIAGNOSTICO - CID-O #3	categórico
Metastase ao DIAGNOSTICO - CID-O #4	categórico
Data do tratamento	datetime
Combinação dos Tratamentos Realizados no Hospital	categórico
Ano do diagnóstico	datetime
Lateralidade do tumor	categórico
Data de Recidiva	datetime
Tempo desde o diagnóstico até a primeira recidiva	quantitativo
Local de Recidiva a distancia/ metastase #1 - CID-O - Topografia	categórico
Local de Recidiva a distancia/ metastase #2 - CID-O - Topografia	categórico
Local de Recidiva a distancia/ metastase #3 - CID-O - Topografia	categórico
Local de Recidiva a distancia/ metastase #4 - CID-O - Topografia	categórico
Descrição da Morfologia de acordo com o CID-O (CID-O - 3ª edição)	categórico
Descrição da Topografia	categórico
Classificação TNM Patológico - N	categórico
Classificação TNM Patológico - T	categórico
Com recidiva à distância	categórico
Com recidiva regional	categórico
Com recidiva local	categórico
TABELA PESO E ALTURA	

nome da coluna	tipo da coluna
Record ID	numérica
Repeat Instrument	numérica
Repeat Instance	numérica
Data:	datetime
Peso	quantitativo
Altura (em centimetros)	quantitativo
IMC	quantitativo

b) Estatística descritiva das colunas.

Foi feita a estatística descritiva após as linhas serem unificadas com base no id único de cada paciente. Nesse sentido, as colunas estão com sufixos relacionados à instância devida, ou seja, '_1' ou '_2'.

O próprio Profile Reports já contém toda a estatística descritiva das colunas, mas foi optado por gerar também manualmente no próprio colab. Nesse contexto, foi dividido entre estatística descritiva para colunas numéricas e categóricas. Para colunas numéricas, apenas utilizamos a função describe() do Pandas, que nos traz informações sobre contagem de valores, média, desvio-padrão, valor máximo, valor mínimo e quartis.

Já para as colunas categóricas, utilizamos a função value_counts() do Pandas, para que ela faça a contagem de todos os valores que aparecem em cada coluna do nosso dataframe. Em relação a gráficos de frequência, é possível visualizar através do Profile Reports, indicado anteriormente.

Como cada tabela possui diversas colunas, será disponibilizado o link para a célula que contém o resultado em relação à estatística descritiva de cada tipo de variável para cada tabela.

TABELA HISTOPATOLOGIA	
Colunas numéricas	clique aqui para acessar a célula no colab.
Colunas categóricas	clique aqui para acessar a célula no colab.

TABELA DEMOGRÁFICO	
Colunas numéricas	clique aqui para acessar a célula no colab.
Colunas categóricas	clique aqui para acessar a célula no colab.

TABELA REGISTRO DE TUMORES	
Colunas numéricas	clique aqui para acessar a célula no colab.
Colunas categóricas	clique aqui para acessar a célula no colab.

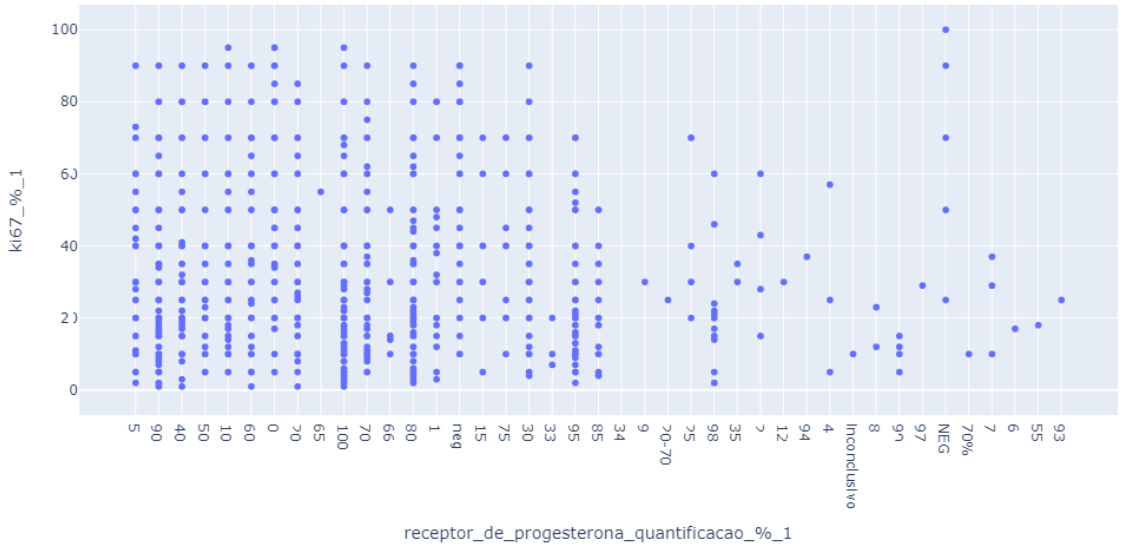
TABELA PESO E ALTURA	
Colunas numéricas	clique aqui para acessar a célula no colab.

c) Relação entre variáveis.

Relação entre variáveis na tabela de histopatologia

I. Relação entre a variável 'Ki67 (%)' e 'Receptorde Estrogênio (quantificação %)'

Figura 08: Relação entre a variável 'Ki67 (%)' e 'Receptorde Estrogênio (quantificação %)'

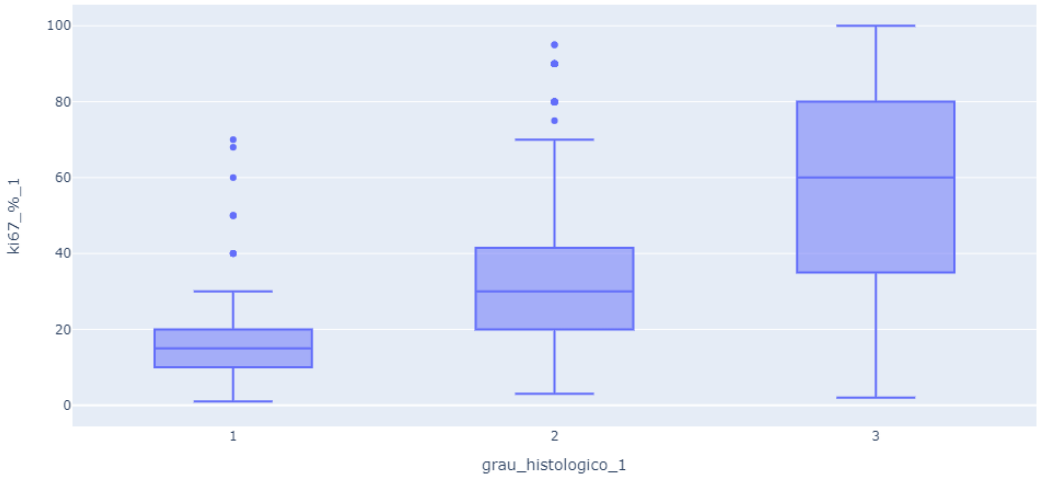


Fonte: Elaboração dos autores.

Percebe-se, através da análise do gráfico acima, a falta de correlação entre estas 2 variáveis, pois os pontos no gráfico estão totalmente dispersos, induzindo que não existe uma correlação clara entre as variáveis.

II. Relação entre a variável 'Grau histológico' e 'Ki67 (%)'

Figura 09: Relação entre a variável 'Grau histológico' e 'Ki67 (%)'



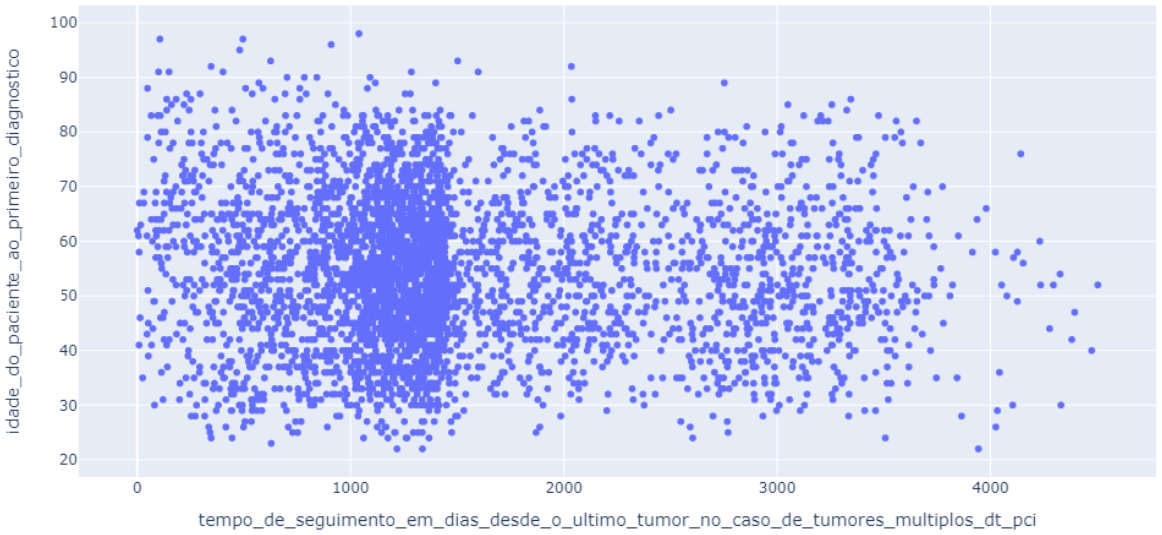
Fonte: Elaboração dos autores.

Já analisando o gráfico anterior, percebe-se claramente uma correlação entre o grau histológico e a quantidade de ki67. Quanto maior o grau histológico, geralmente os pacientes, em média, possuem uma quantidade maior de ki67(%).

Relação entre variáveis na tabela de demográfico

I. Relação entre a variável 'tempo_de_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_de_tumores_multiplos_dt_pci' e 'idade_do_paciente_ao_primeiro_diagnostico'

Figura 10: Relação entre a variável 'tempo_de_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_de_tumores_multiplos_dt_pci' e 'idade_do_paciente_ao_primeiro_diagnostico'

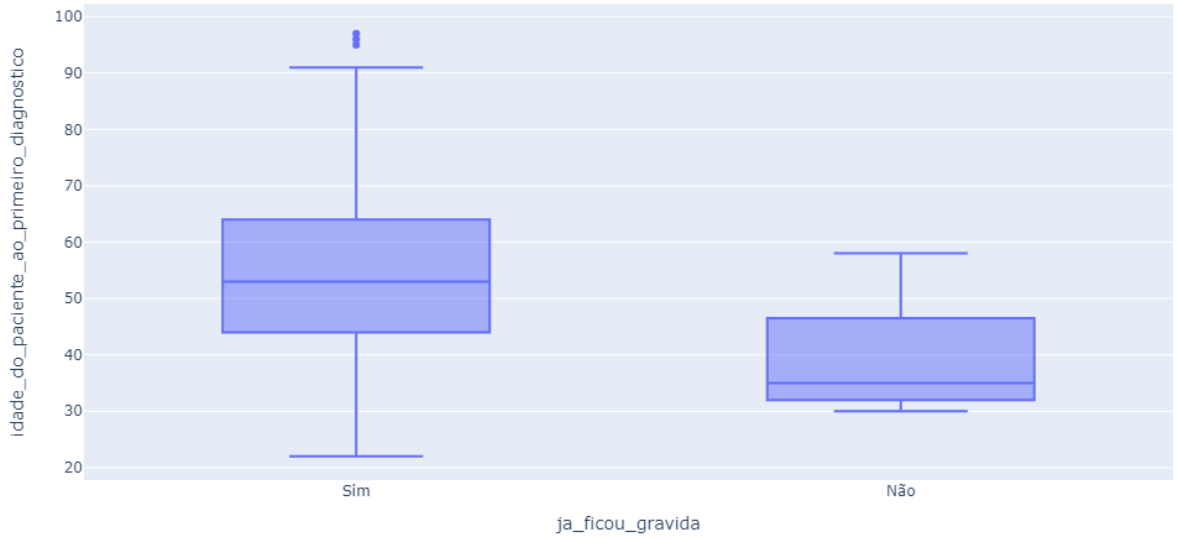


Fonte: Elaboração dos autores.

Percebe-se, através da análise do gráfico anterior, a relação não clara entre estas 2 variáveis, pois os pontos no gráfico estão totalmente dispersos, induzindo que não existe uma correlação clara entre as variáveis, a não ser uma concentração massiva de dados entre o range [1000,1500] no tempo de seguimento e no range [40, 70] na idade do paciente.

II. Relação entre a variável 'idade_do_paciente_ao_primeiro_diagnostico' e 'ja_ficou_gravida'

Figura 11: Relação entre a variável 'idade_do_paciente_ao_primeiro_diagnostico' e 'ja_ficou_gravida'



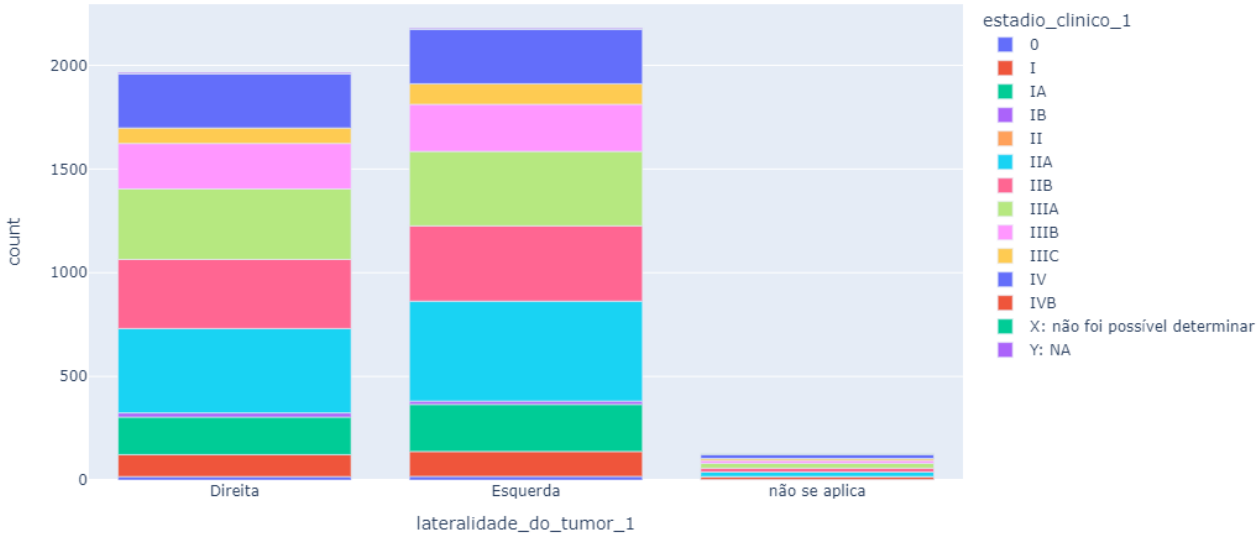
Fonte: Elaboração dos autores.

Já analisando o gráfico anterior, percebe-se que em casos que a paciente já tenha ficado grávida, geralmente, em média, ela possui uma idade maior do que se não tivesse ficado grávida. Outro ponto interessante é a concentração entre 30 e 60 anos para mulheres que nunca ficaram grávidas.

Relação entre variáveis na tabela de histopatologia

I. Relação entre a variável 'estadio_clinico' e 'lateralidade_do_tumor'

Figura 12: Relação entre a variável 'estadio_clinico' e 'lateralidade_do_tumor'

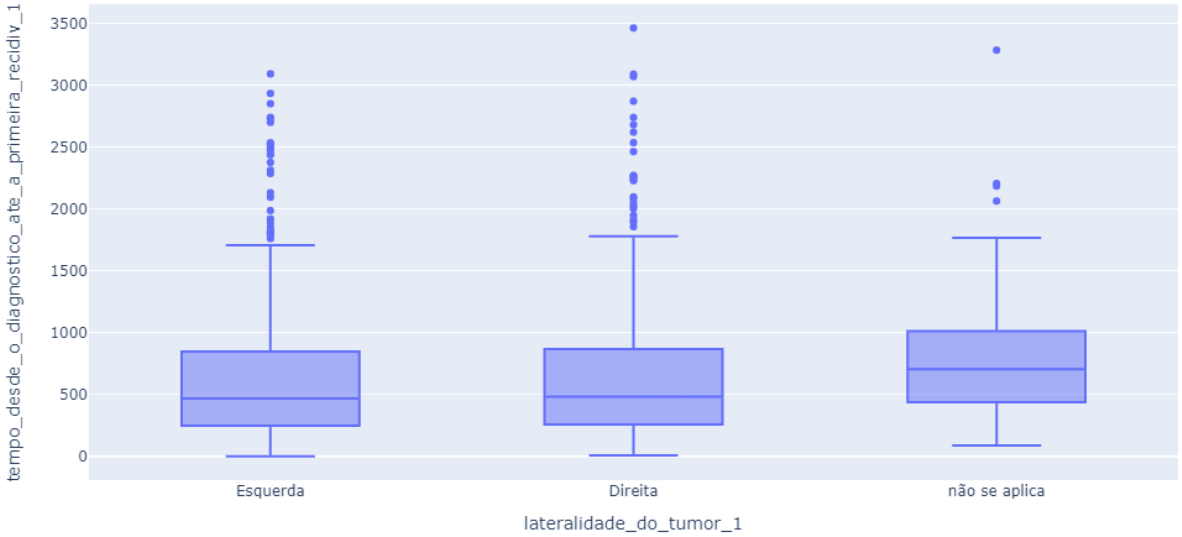


Fonte: Elaboração dos autores.

Através do gráfico anterior é possível perceber que a porporção da contagem de cada estadio clinico é relativamente semelhante para pacientes que possuem lateralidade na parte direita ou esquerda, onde apenas no estadio clinico IIA a proporção de pacientes que tiveram lateralidade do tumor na parte esquerda é maior.

II. Relação entre a variável 'lateralidade_do_tumor_1' e 'tempo_desde_o_diagnostico_ate_a_primeira_recidiv_1'

Figura 13: Relação entre a variável 'lateralidade_do_tumor_1' e 'tempo_desde_o_diagnostico_ate_a_primeira_recidiv_1'



Fonte: Elaboração dos autores.

O gráfico anterior apresenta uma alta semelhança entre os tipos de lateralidade do tumor e o tempo do diagnóstico até a primeira recidiva.

2. Pré-processamento dos dados:

A etapa de pré-processamento dos dados é um conjunto de técnicas aplicadas aos dados brutos coletados, para que possam ser adequadamente utilizados em algoritmos de aprendizado de máquina e outras técnicas de modelagem. Nesse sentido, ocorre a limpeza de dados, tratamento de dados faltantes e tratamento dos outliers. Além disso, ocorre também a padronização dos dados, através de encoding das variáveis categóricas e a normalização dos dados numéricos. Essa etapa é de suma importância pois ajuda a economizar recursos computacionais, melhorando, também, a eficácia dos modelos preditivos.

Nesse contexto, essa etapa foi subdividida em exclusão de colunas desnecessárias, agrupamento das colunas, tratamento dos dados faltantes, tratamento dos outliers, encode das variáveis categóricas e normalização das variáveis numéricas.

Exclusão de colunas

Nessa etapa será excluído colunas consideradas irrelevantes para o nosso modelo, seja por não ter dados úteis (dados iguais) ou por não ter dados suficientes. Dessa forma, foi criado 2 funções: 'delete_columns_intances()' que recebe como parâmetros 'df' e 'columns' e a função 'delete_columns', que recebe os mesmos parâmetros. A diferença entre as 2 funções é que a primeira é relacionada a colunas com instâncias (que possuem sufixo _1 e _2) e a segunda são colunas sem instâncias. Para acessar no colab: clique aqui.

Agrupamento das colunas

Nessa etapa foi realizado o agrupamento de colunas semelhantes, a fim de diminuir a dimensionalidade da nossa base, além de criar features mais "potentes". Além disso, é de suma importância realizar o agrupamento das colunas antes do tratamento dos dados faltantes, pois muitas vezes a informação está dispersa em mais de 1 coluna, onde, ao agrupar, é possível obter a informação completa, sem que seja necessário a imputação de dados, ou pelo menos reduzindo essa imputação. Dessa forma, a escolha das colunas para agrupamento foi ao encontro com as atuações descritas no Sheets explicitado na seção de análise exploratória. Para acessar no colab: clique aqui.

Missings

Missings são registros ausentes em um dataset, ou seja, são dados não informados ou não preenchidos. Estes dados são importantes para análises e inferências estatísticas, pois são inseridos valores ausentes que podem influenciar nos resultados.

A etapa de tratamento de dados faltantes é extremamente importante, pois não se pode imputar dados nulos nos modelos. Além disso, é importante para garantir que o modelo seja treinado com precisão e eficiência, evitando resultados imprecisos e não confiáveis. Dessa forma, foi optado por imputar dados via distribuição normal, a fim de manter a proporcionalidade e o comportamento dos dados dentro de cada coluna, visto que ao imputar a média, mediana ou moda pode acabar gerando um viés. Para acessar no colab: clique aqui.

Outliers

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

   Os outliers são basicamente valores que estão muito acima ou muito abaixo de todos os valores de determinado conjunto de dados, isso significa que estes valores considerados outliers podem exercer uma influência desproporcional sobre a média, puxando a muito para cima ou muito para baixo, levando-a a se desviar

significativamente do valor que seria observado se esses outliers não estivessem presentes.

Identificar e tratar esses outliers é extremamente importante para o modelo preditivo, pois eles podem enviesar o modelo, já que o modelo pode considerar esses outliers como um padrão significativo, e acabar dando mais peso a eles.

Para tratar esses outliers, criamos uma função que identifica eles com base no desvio-padrão e na média. Isso foi feito calculando um intervalo, multiplicando o desvio-padrão por 2,7. Após calcular esse intervalo, determinamos o limite superior e inferior.

O limite superior foi determinado somando a média e intervalo calculado. Já o limite inferior foi determinado subtraindo a média do intervalo calculado.

Após o cálculo do intervalo e a determinação do limite inferior e superior, criamos um dataframe à parte que mostra o nome da coluna onde os outliers foram encontrados e os outliers encontrados acima do limite superior e inferior:

	col	lower_outliers	upper_outliers
0	record_id	[]	[]
1	repeat_instance_x	[]	[3.0, 3.0, 3.0, 3.0, 3.0, 3.0, 4.0, 5.0, 6.0, ...
2	grau_histologico	[]	[]
3	subtipo_tumoral	[]	[]
4	indice_h_(receptorde_progesterona)	[]	[]
5	ki67_(%)	[]	[]
6	repeat_instance_y	[]	[3.0, 3.0, 3.0, 3.0, 3.0, 4.0, 5.0, 6.0, 3.0, ...
7	codigo_da_morfologia_de_acordo_com_o_cido	[81403.0, 81403.0, 80903.0, 80103.0, 80973.0, ...	[99873.0, 88903.0, 89803.0, 88013.0, 96803.0, ...
8	ano_do_diagnostico	[]	[]
9	tempo_desde_o_diagnostico_até_a_primeira_recid...	[]	[2442.0, 2739.0, 2184.0, 2437.0, 2534.0, 2977....
10	repeat_instrument	[]	[]
11	repeat_instance	[]	[]
12	idade_do_paciente_ao_primeiro_diagnostico	[]	[93.0, 92.0, 93.0, 92.0, 97.0, 95.0, 95.0, 97....
13	tempo_de_seguimento_(em_dias)__desde_o_último_...	[]	[4153.0, 4330.0, 4381.0, 3734.0, 4474.0, 3864....
14	quantas_vezes_ficou_gráfida	[]	[7.0]
15	número_de_partos	[]	[]
16	idade_na_primeira_gestacao	[0.0, 0.0, 0.0, 0.0]	[39.0, 45.0, 40.0, 42.0, 39.0, 41.0, 42.0, 53....
17	por_quanto_tempo_amamentou	[]	[82.0, 100.0, 178.0, 84.0, 96.0, 240.0, 150.0,...
18	idade_da_primeira_mentruacao	[0.0, 7.0]	[37.0, 19.0, 30.0, 19.0, 20.0]

Com os outliers identificados, fizemos uma outra função que remove esses outliers. Para essa função, fizemos os mesmos cálculos mostrados anteriormente. Porém, a única diferença foi que ao invés de criarmos um dataframe à parte, utilizamos a função drop() do Python para remover esses valores de cada coluna.

Por fim, as colunas que encontramos outliers foram as seguintes:

['tempo_desde_o_diagnostico_até_a_primeira_recidiva__', 'idade_do_paciente_ao_primeiro_diagnostico', 'tempo_de_seguimento_(em_dias)__desde_o_último_tumor_no_caso_de_tumores_múltiplos____[dt_pci]', 'quantas_vezes_ficou_gráfida', 'idade_na_primeira_gestacao', 'por_quanto_tempo_amamentou', 'idade_da_primeira_mentruacao']

Para acessar o colab, clique aqui.

3. Hipóteses:

Nessa etapa, é preciso formular hipóteses sobre as relações entre as variáveis que serão usadas no modelo. Isso ajuda a direcionar a análise dos dados e a construção do modelo, identificando as variáveis mais importantes e as técnicas estatísticas adequadas

Com isso, foram criadas 6 hipóteses iniciais para o projeto:

1. Para casos em que há metástase, é mais indicada a terapia neoadjuvante;
2. Para mulheres que já ficaram grávidas, o tratamento mais indicado é o adjuvante;
3. Para mulheres jovens (20-30 anos) o tratamento mais indicado é o tratamento adjuvante.
4. Caso a pessoa comece o tratamento indicado em até 2 meses após o diagnóstico do câncer de mama, a probabilidade de sucesso é superior a 80%.
5. Para casos de quantidade de progesterona superior a 70%, o tratamento mais indicado é o tratamento neoadjuvante.
6. Para mulheres já caracterizadas na menopausa (50+ anos), o tratamento mais indicado é o neoadjuvante.

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

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

Um exemplo de referência de livro:

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

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

Anexos

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