



Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
01/02/2023	Yago Phellipe	0.1	Preenchimento da seção 4.1.3.
02/02/2023	José Vitor Marcelino	0.2	Implementação de Personas 4.1.6
06/02/2023	José Vitor Marcelino	0.3	Rascunho da Introdução 1.0 (Pré-Validação com a professora)
06/02/2023	Emely Vitória	0.4	Preenchimento da seção 4.1.1
07/02/2023	Marcos Teixeira	0.5	Preenchimento da seção 4.1.5
08/02/2023	Marcos Teixeira	0.6	Preenchimento da seção 4.1.5
08/02/2023	Yuri Toledo	0.7	Preenchimento da seção 4.1.7
09/02/2023	Marcos Teixeira	0.8	Preenchimento da seção 4.1.3
10/02/2023	Emely Vitória	0.9	Preenchimento da seção 2.1
20/02/2023	Yuri Toledo	1.1	Correção dos erros apontados pela professora na última sprint
23/02/2023	Vivian Shibata	1.2	Preenchimento da LGPD, análise SWOT e ABNT
26/02/2023	Vivian Shibata	1.3	Preenchimento das hipóteses na compreensão dos dados
26/02/2023	Yago Phellipe	1.4	Preenchimento da questão 1 letra A e B da seção 4.2
26/02/2023	Emely Vitória	1.5	Preenchimento da seção 4.2.1.2 letra A
27/02/2023	Daniel Dávila	1.6	Reescrita dos itens 1 e 2; preenchimento do item 3

07/03/2023	José Vitor Marcelino, Vivian Shibata	1.7	Reescrita do item 3 e preenchimento da seção 4.3 letra A
09/03/2023	Yuri Toledo	1.8	Preenchimento da letra b da seção 4.3
11/03/2023	Yuri Toledo	1.9	Preenchimento da letra c da seção 4.3
03/04/2023	Daniel Dávila	2.0	Completa reescrita da seção 4, incluindo criação das subseções 4.3.3., 4.4.1, 4.4.2, e 4.4.3.
04/04/2023	José Vitor Marcelino, Vivian Shibata	2.1	e Revisão da 4.4.2

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. Cinco forças de Porter	7
4.1.3. Análise SWOT	7
4.1.4. Planejamento Geral da Solução	7
4.1.5. Value Proposition Canvas	7
4.1.6. Matriz de Riscos	7
4.1.7. Personas	8
4.1.8. Jornadas do Usuário	8
4.1.9 Política de privacidade para o projeto de acordo com a LGPD	8
4.2. Compreensão dos Dados	9
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

O Instituto do Câncer de São Paulo (ICESP) é um dos maiores hospitais especializados no tratamento de câncer da América Latina. Sua sede é um prédio com mais de 100 metros de altura e 80 mil metros quadrados, localizado na Avenida Doutor Arnaldo, na zona oeste da cidade de São Paulo.

O instituto surgiu em 2008 (embora o edifício já exista há mais de 30 anos) como resultado de uma parceria entre a Faculdade de Medicina da USP (FMUSP) e o Governo de São Paulo. No começo da elaboração de seu projeto, em 1987, o instituto inicialmente tinha a finalidade de ser um centro médico com programas focados na área da saúde da mulher. Anos antes de sua inauguração, em 2005, o Conselho Deliberativo do HCFMUSP (Hospital das Clínicas da Faculdade de Medicina da USP) apresentou uma proposta que tornaria o projeto não só uma unidade exclusiva para saúde feminina, mas também um hospital que realizaria transplantes, cirurgias complexas e oncologia cirúrgica. Ao contrário do previsto, em dezembro de 2007, por uma decisão do governador de São Paulo da época, o hospital se tornou uma unidade totalmente dedicada ao paciente oncológico.

O instituto tem como principal área de atuação o tratamento de câncer, já que seu principal propósito é ser referência de assistência, ensino e pesquisa em oncologia. Inclusive dentro do ICESSP existe o CTO (Centro de Investigação Translacional em Oncologia), que possui vários grupos de pesquisadores que investigam novos medicamentos e métodos de tratamento.

Por ser uma organização pública, o ICESSP não possui rivais de mercado, já que seus serviços possuem um viés colaborativo/educativo e suas pesquisas podem ser utilizadas por outras instituições da área que estejam interessadas em contribuir com a ciência e a sociedade. Entretanto, ocasionalmente o hospital presta serviços médicos pagos para complementar custos operacionais. Existem outros hospitais que também trabalham na área da oncologia e todos, incluindo o ICESSP, estão investindo bastante em tecnologia e em processos inovadores, como, por exemplo, o uso das análises diagnóstica, preditiva e prescritiva de dados.

Devido ao seu foco em oncologia, o ICESSP enxerga que a evolução do câncer de mama ainda é algo bastante variável durante os tratamentos oncológicos convencionais. Sendo assim, o ICESSP deseja descobrir um padrão preditivo existente entre os pacientes diagnosticados com câncer para saber qual tipo de tratamento é o melhor em cada caso: Neoadjuvante (1º quimioterapia e 2º cirurgia) ou Adjuvante (1º cirurgia e 2º quimioterapia).

2. Objetivos e Justificativa

2.1. Objetivos

O principal objetivo do parceiro de negócio é criar um modelo preditivo capaz de auxiliar o corpo médico na tomada de decisão sobre o tratamento de uma paciente. Para isso, serão realizadas as seguintes tarefas:

1. Analisar uma ampla variedade de informações, incluindo dados clínicos (informações sobre saúde, histórico médico e medicações) e demografia (informações sobre idade, gênero, renda, educação, etc).
2. Filtrar as informações essenciais do paciente para prever uma saída futura com base em dados e relações entre as variáveis de entrada.
3. Classificar dados de pacientes com câncer a fim de identificar qual é a melhor forma de realizar o tratamento de câncer de mama: neo (1º quimioterapia e 2º cirurgia) ou adjuvante (1º cirurgia e 2º terapia), para, assim, detectar padrões que indiquem aos profissionais de saúde uma possível rota de tratamento indicado para cada perfil dos pacientes.

2.2. Proposta de Solução

A equipe desenvolverá um modelo preditivo que tem como objetivo recomendar o tratamento adequado para cada caso específico. O produto vai focar em resolver o problema da variabilidade de respostas a tratamentos de câncer de mama, prevendo qual dos dois tratamentos (adjuvante e neoadjuvante) é o melhor para cada paciente, auxiliando o médico na escolha do tratamento apropriado. O modelo terá os dados de pacientes diagnosticados com câncer fornecidos pelo parceiro como base, esses dados servirão para análise e treinamento do algoritmo do modelo.

2.3. Justificativa

O tratamento convencional do câncer de mama possui resultados muito variados, o que certamente atrapalha qualquer conclusão médica na hora da sugestão de qual tratamento deve ser o adequado. A ordem dos processos médicos (quimioterapia e cirurgia) aplicados durante o tratamento de um paciente tem influência direta na taxa de sucesso de remoção do câncer, sendo assim, muito importante definir qual ordem é a mais apropriada a seguir.

Tendo esse problema em vista, um modelo preditivo baseado na análise de dados clínico-laboratoriais é uma solução que pode auxiliar o médico e sustentar, por meio de evidências, qual tratamento (adjuvante ou neoadjuvante) é o melhor para o caso de cada paciente.

3. Metodologia

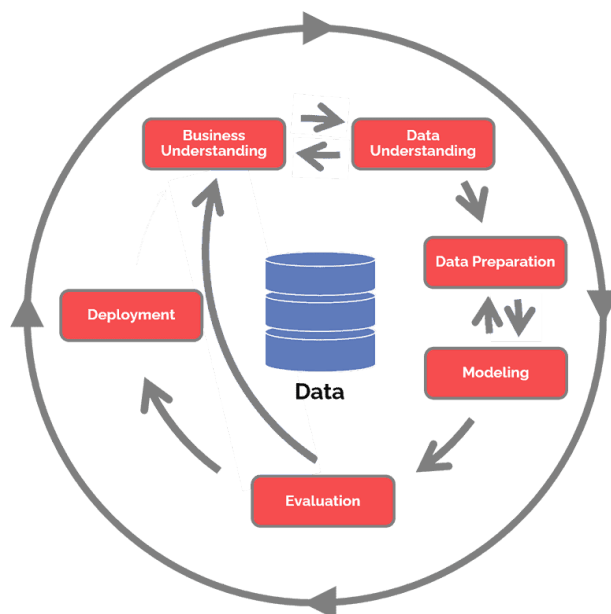


Figura 1 - Metodologia CRISP-DM

O CRISP-DM (Cross Industry Standard Process for Mining Data) é uma metodologia que consiste em um conjunto de práticas para mineração de dados. Esse método possui 6 passos bem definidos, sendo eles: Entendimento do negócio, Entendimento dos dados, Preparação dos dados, Modelagem dos dados, Avaliação do modelo e Deployment.

1º Passo - Entendimento do negócio: O entendimento do negócio consiste em compreender o problema a ser resolvido e o objetivo do projeto, sempre evitando vieses inconscientes. É importante captar todos os detalhes não só do problema e do projeto, mas também da empresa para entender sua estratégia.

2º Passo - Entendimento dos dados: O entendimento dos dados é o passo no qual devemos organizar, documentar e conhecer os dados. É extremamente necessário compreender a fonte dos dados, verificar a qualidade deles e estudá-los de maneira minuciosa para distinguir quais dados são relevantes para o projeto e quais podem ser descartados.

3º Passo - Preparação dos dados: Essa etapa consiste basicamente no pré-processamento dos dados, devemos tratar os dados conforme nosso interesse e

necessidade, excluir anomalias e dados vazios, normalizar e padronizar dados. Essa é a fase mais complexa da CRISP-DM, demandando cerca de 70%-90% do tempo do projeto, portanto é crucial fazer essa parte muito bem feita, para que não seja necessário ficar voltando constantemente nessa fase.

4º Passo - Modelagem: Na etapa da Modelagem escolhemos o tipo de modelagem ideal para a nossa base de dados por meio de ferramentas computacionais com o objetivo de resolver o problema analisado na primeira etapa (Entendimento do negócio).

5º Passo - Avaliação do modelo: Na avaliação conseguimos analisar e verificar se o resultado do modelo corresponde a nossa expectativa e é satisfatório para a equipe. Em caso negativo é recomendável reavaliar as etapas anteriores em busca de aprimorar o modelo preditivo.

6º Passo - Deployment: A última etapa do CRISP-DM é o Deployment que basicamente é colocar o modelo em ação de modo a agregar valor para o negócio. O modelo costuma ficar armazenado na nuvem ou em servidores locais do cliente.

Em algumas situações do diagrama, pode ser necessário retornar a etapas anteriores por conta de falhas durante o processo ou entendimento insuficiente do negócio, essas voltas são importantes para aprimorar o processo de exploração de análise dos dados.

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

Após uma análise de mercado, não foram identificadas empresas que possam ser consideradas concorrentes diretas do parceiro de negócios, já que ele se posiciona como uma instituição pública dedicada a servir a sociedade. Todo o conhecimento produzido internamente pode ser usado por outras organizações, desde que sejam respeitadas as regras de uso e direitos comerciais. No entanto, existem outras empresas que atuam na mesma área, como o Hospital A.C.Camargo, especializado no tratamento e pesquisa do câncer(CA); O Centro oncológico Família Dayan - Daycoval, pertence ao Hospital Israelita Albert Einstein, e o INCA, um órgão vinculado ao Ministério da Saúde responsável pela prevenção e controle do câncer no Brasil.

O Hospital das Clínicas é uma instituição de saúde pública universitária, localizada em São Paulo, Brasil, que possui um setor interno para o tratamento de CA, o Instituto do Câncer do Estado de São Paulo (ICESP). Como uma instituição pública, o hospital não tem um modelo de

negócios tradicional baseado em lucro, pois é financiado pelo governo e pela universidade para fornecer atendimento médico gratuito à população. Além disso, o Hospital das Clínicas também oferece serviços médicos pagos, como exames e procedimentos, para aqueles que têm capacidade financeira para pagá-los, o que ajuda a complementar o financiamento da instituição. Em resumo, o modelo de negócios do Hospital das Clínicas é misto, incluindo financiamento público e serviços médicos pagos para complementar sua operação.

Considerando o ambiente de mercado em que o hospital está inserido, pode-se ver que ele está sempre em busca de melhorias, investindo em modernização, inovação tecnológica e plataforma educacional, além de firmar parcerias com a iniciativa privada e expandir o processo de internacionalização para ser um centro educacional de referência no mundo (USP, 2021). Essas inovações incluem a procura por equipamentos mais avançados, tratamentos mais eficazes e menos prejudiciais, entre outros aspectos.

Abaixo uma análise de indústria utilizando as cinco forças de Porter.

Rivalidade entre os concorrentes: Por ser uma instituição governamental sem fins lucrativos, o ICESP não tem concorrentes diretos, mas sim hospitais parceiros com os quais trabalha na pesquisa e no tratamento do CA. No entanto, existem outras instituições, como o Instituto Nacional do Câncer, e vários hospitais privados que também oferecem tratamento para o câncer e competem por pacientes.

Poder de barganha de clientes: É variável, dependendo do nível socioeconômico e da disponibilidade de opções de saúde para eles. No entanto, a instituição possui uma grande demanda por parte da população, o que pode limitar o poder de negociação dos clientes.

Poder de barganha de fornecedores: O parceiro precisa de produtos muito avançados e de elevado valor agregado, o que significa que seus fornecedores têm grande poder de barganha. Isso acontece porque certos medicamentos e equipamentos de pesquisa não são amplamente disponíveis e sua produção e preço são controlados por um pequeno grupo de empresas. Como fornecedores de dados, que são a base para as atividades do parceiro, os hospitais também podem recusar fornecê-los. Além disso, as tecnologias criadas requerem um longo processo de licenciamento, e os responsáveis pelo desenvolvimento podem fazer exigências quanto ao custo da solução.

Ameaças de produtos substitutos: É moderado devido ao fato de que há cada vez mais investimentos da iniciativa privada no desenvolvimento de novas tecnologias e serviços, podendo haver novas soluções e tratamentos mais eficazes.

Ameaças de novos entrantes: Não há preocupação com a chegada de novos competidores, pois as instituições na mesma área de atuação do Hospital das Clínicas não são vistas como concorrentes, mas sim como parceiros. A FMUSP é reconhecida como uma referência na área da

saúde, e qualquer nova empresa que surja no mercado não seria capaz de competir com ela a curto prazo.

4.1.2. Análise SWOT

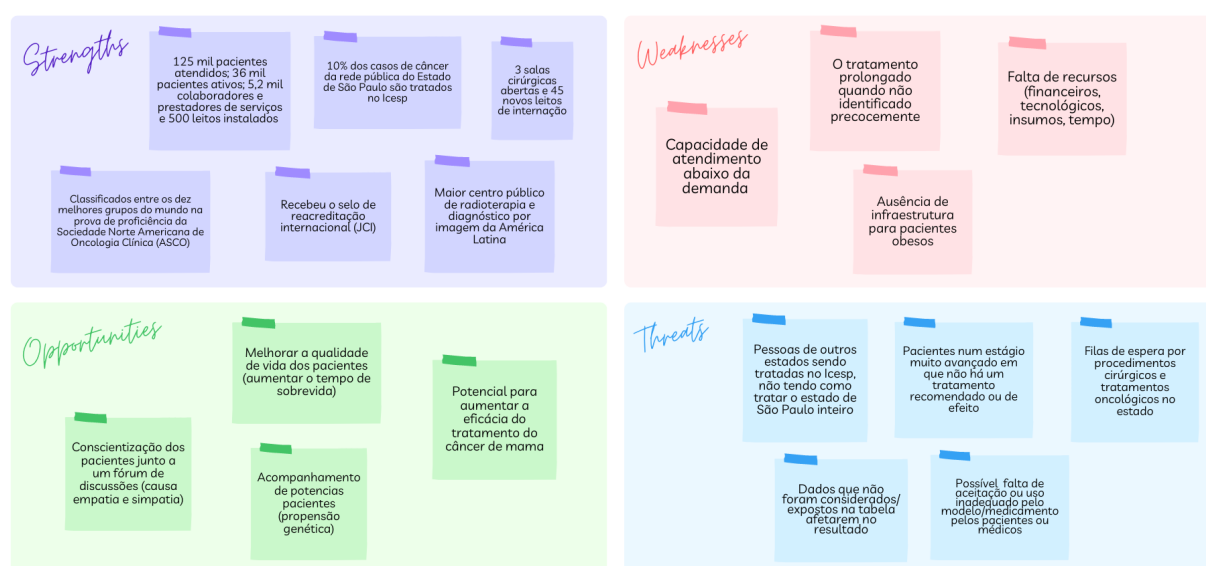


Figura 2 - Matriz SWOT

Fonte: Elaboração própria

(Forças)

De acordo com a plataforma oficial do Instituto do Câncer do Estado de São Paulo (ICESP, 2022), - é responsável por cerca de 10% dos casos de câncer da rede pública do Estado de São Paulo - desde a sua inauguração em 6 de maio de 2008, o instituto atendeu aproximadamente 125 mil pacientes, com 36 mil pacientes ativos e contam com 5,2 mil colaboradores e prestadores de serviços. Além disso, há 445 leitos operacionais e na sua capacidade máxima disponibilizará 490 leitos, sendo 85 de Unidade de Terapia Intensiva (UTI).

Recentemente, houve uma ampliação em sua infraestrutura onde foram abertas mais três salas cirúrgicas e 45 novos leitos de internação, sendo 15 são de UTI. Isso possibilitará o atendimento de 1.250 novos pacientes no próximo ano e a realização de 840 cirurgias adicionais, o que seria o equivalente a um incremento de 20% do previsto para o período. A

iniciativa visa reduzir a fila atual de pacientes oncológicos no Estado de São Paulo em 40%, nos primeiros meses da ação (ICESP, 2023).

O ICESSP foi o primeiro hospital da rede pública da capital a receber o selo de reacreditação pela Joint Commission International (JCI) - uma certificação internacional que reconhece a excelência no atendimento e serviços oferecidos à população - em 2014. O selo passa por um processo de verificação que ocorre de três em três anos, o instituto se submete à avaliação a fim de manter o selo, a última verificação ocorreu em 2020 (GOVERNO, 2021).

Sobre a qualidade do ensino e pesquisa realizados, o ICESSP se destaca no Programa de Residência Médica em Cancerologia Clínica, originado em 1998 como primeiro do país. Desde sua criação, já foram formados mais de 170 médicos oncologistas e atualmente é um dos maiores programas do país, reconhecido nacional e internacionalmente.

Nos últimos anos, os formandos foram sistematicamente classificados entre os dez melhores grupos do mundo na prova de proficiência da Sociedade Norte Americana de Oncologia Clínica (ASCO). Porém, no ano de 2022, os residentes do segundo e terceiro ano do programa de residência médica de Oncologia Clínica do ICESSP, alcançaram o melhor desempenho no exame anual da Sociedade Americana de Oncologia Clínica (ASCO). O grupo obteve a maior média entre todas as instituições avaliadas e atingiu sua maior pontuação em relação a avaliações de anos anteriores, superando assim sua melhor marca. O que coloca a instituição no topo dos melhores profissionais de oncologia do mundo (ICESP, 2023).

(Fraquezas)

Um grande agravante na falha do processo do tratamento e na alta demanda, seria a falta de recursos. Por exemplo, a imunoterapia pode trazer benefícios como a alta eficácia e baixos efeitos colaterais, no entanto, uma única caixa desse medicamento pode custar 15 mil reais. Contudo, no Sistema Único de Saúde (SUS) são poucas as novas drogas que são oferecidas, essa dificuldade de acesso frequentemente leva a algumas pessoas a até entrarem na justiça para a obtenção do tratamento.

Segundo Maria Del Pilar Estevez Diz, diretora do Corpo Clínico do ICESSP, diz que a dificuldade de acesso ao uso de novas tecnologias é uma das principais razões para que o câncer de mama tenha ainda alta mortalidade no Brasil. "A maior parte dos médicos atua no SUS e no setor privado e, com isso, vivencia situações muito díspares. Falta equidade." Como há uma alta demanda pelos tratamentos, os médicos não têm tempo suficiente para propriamente entender o paciente num nível mais pessoal, o que faz com que a atenção e o atendimento não sejam tão precisos.

O oncologista Stephen Stefani, presidente da *International Society for Pharma-coeconomics and Outcomes research* (Ispor) no Brasil, relata que uma das origens do

problema de acesso são as distorções no sistema. “Apenas 25% das pessoas no Brasil têm acesso a planos de saúde, mas 55% dos recursos no País são gastos com essa população”, afirmou. Além disso, há novos tratamentos contra o câncer que podem custar até US\$ 10 mil quando chegam ao mercado – e pesam no sistema, aumentando a distorção. “Qualquer incorporação de um novo medicamento, se não for feita com cuidado, pode aumentar o número de excluídos. Os recursos são limitados, e não podemos conceder qualquer tipo de desperdício.” (O ESTADO DE S.PAULO, 2019).

(Ameaças)

O Ministério Público Federal (MPF) entrou com uma ação contra o Governo de São Paulo para que providências sejam tomadas a respeito da lei federal que determina que pacientes com câncer devem receber tratamento em até 60 dias após o diagnóstico. O MPF destacou que mais de 18 mil pessoas aguardam mais de dois meses entre o diagnóstico e o começo da terapia.

De acordo com a Secretaria Estadual da Saúde (SES), 1.536 pessoas continuam na espera por cirurgias para tratamento de câncer no estado e, em alguns casos, a espera chega a ser de oito meses (G1 SP, 2023).

Para poder ser atendido no ICESP, o paciente precisa ser de uma certa localidade para poder ser atendido, mas há casos em que alguns pacientes trocam o seu comprovante de residência para poder receber o tratamento.

A boa alimentação é um fator importante para a prevenção do câncer, manter uma dieta equilibrada pode ajudar na prevenção de diversas doenças. Alimentos ricos em fibras, vitaminas e antioxidantes oferecem inúmeros benefícios ao organismo. Portanto, a má alimentação da população pode agravar os casos de câncer (ICESP, 2022).

(Oportunidades)

A Organização Mundial da Saúde (OMS) recomenda ações de prevenção, detecção precoce e acesso ao tratamento para controle melhor do câncer, pois quanto mais cedo o câncer for identificado, maiores são as chances de cura. A detecção precoce do câncer consiste em duas estratégias: a primeira seria o rastreamento, que tem por objetivo encontrar o câncer pré-clínico ou as lesões pré-cancerígenas, por meio de exames de rotina em uma população-alvo sem sinais e sintomas sugestivos do câncer rastreado. A segunda, corresponde ao diagnóstico precoce, que busca identificar o câncer em estágio inicial em pessoas que apresentam sinais e sintomas suspeitos da doença (INCA, 2021).

A conscientização da população também é um fator importante para a prevenção do câncer. Por exemplo, segundo estatísticas do Instituto Nacional de Câncer (INCA), o tabagismo é a principal causa de câncer de pulmão evitável no mundo e as consequências da queima do cigarro são sentidas não apenas por quem fuma, mas também por todos ao seu redor. Outro exemplo seria o consumo de álcool que, de acordo com a Agência Internacional de Pesquisa sobre o Câncer, a quantia de 18 gramas (aproximadamente duas doses) de álcool por dia era suficiente para aumentar significativamente o risco de desenvolver câncer de mama. Consequentemente, com pessoas mais bem conscientizadas e proativas sobre o assunto, o número de pacientes diminui (GIGLIO, 2021).

4.1.3. Planejamento Geral da Solução

a) Qual é o problema a ser resolvido?

A maior dificuldade que médicos de câncer de mama enfrentam atualmente consiste em decidir qual é o tratamento ideal para cada paciente, posto que o método varia entre cada indivíduo. Dessa forma, desprovidos de uma métrica acurada, os médicos tendem a gastar muito tempo analisando os dados de cada paciente para tentar ver o padrão daquela pessoa que, por ser um processo manual, ainda está sujeito a erros, podendo ser indicado o tratamento que não seria tão eficaz para aquele indivíduo. E, desse modo, precisar de um modelo preditivo para responder com certeza aos questionamentos: "Realizar a cirurgia primeiro e a quimioterapia depois?"; "Realizar a quimioterapia primeiro e a cirurgia depois?".

b) Qual a solução proposta (Visão de negócios).

Desenvolver um modelo preditivo para ajudar médicos a escolher o melhor tratamento para seus pacientes com câncer de mama. A plataforma coleta e analisa múltiplos dados de pacientes passados, incluindo informações clínicas e resultados de tratamentos, para criar um modelo preditivo que indica a probabilidade de sucesso de cada tratamento para um paciente específico.

Esta solução resolve o dilema que muitos médicos enfrentam ao tentar decidir qual tratamento é melhor para cada paciente, pois fornece uma base sólida de dados e análise para apoiar suas decisões clínicas. Além disso, a plataforma pode ajudar a garantir que os pacientes recebam o tratamento mais eficaz e aumentar a eficiência do sistema de saúde, pois permite que os médicos tomem decisões mais informadas e baseadas em evidências.

A partir de uma perspectiva de negócios, esta solução pode se destacar em um mercado em constante evolução e com crescente demanda por soluções tecnológicas avançadas na

medicina. Além disso, a plataforma pode ser oferecida como um serviço a hospitais, clínicas e grupos médicos, dispostos a pagar pela realização do serviço, gerando um grande lucro.

c) Qual o tipo de tarefa (regressão ou classificação).

O nosso modelo é de classificação, pois ele tem como target o domínio 0 ou 1, que significa insucesso ou sucesso, respectivamente. Apesar de apresentar uma lógica de quantificação, nós retornamos esse score em um valor binário de 0 ou 1, onde se o score for maior ou igual a 0 é tido como sucesso e caso contrário é insucesso. Dessa forma, nosso modelo se enquadra no molde de classificação.

d) Como a solução proposta deverá ser utilizada.

A solução proposta deverá ser utilizada por um médico especialista no tratamento do paciente, ciente de que o modelo preditivo trata-se de, no máximo, uma *recomendação*. A decisão final sobre o melhor tratamento deve invariavelmente basear-se na análise dos dados coletados pelo médico sobre determinado paciente.

e) Quais os benefícios trazidos pela solução proposta.

O dilema que muitos médicos enfrentam ao tentar decidir qual tratamento é ideal para cada paciente possui sua resolução extremamente facilitada por meio da utilização do modelo. Dessa forma, o modelo aumenta as chances de que pacientes recebam o tratamento mais eficaz, e, ao permitir que médicos tomem decisões mais informadas e mais embasadas mais rapidamente, aumenta a eficiência do sistema como um todo.

f) Qual será o critério de sucesso e qual métrica será utilizada para avaliá-lo.

O critério de sucesso é o aumento do tempo de sobrevida dos pacientes, tendo como métricas o fato de a paciente ser reincidente e/ou se houve extensão no tempo estimado de sobrevida. Ressaltando-se que a extensão do tempo de sobrevida é o principal fator analisado no entendimento do sucesso do modelo.

4.1.4. Value Proposition Canvas

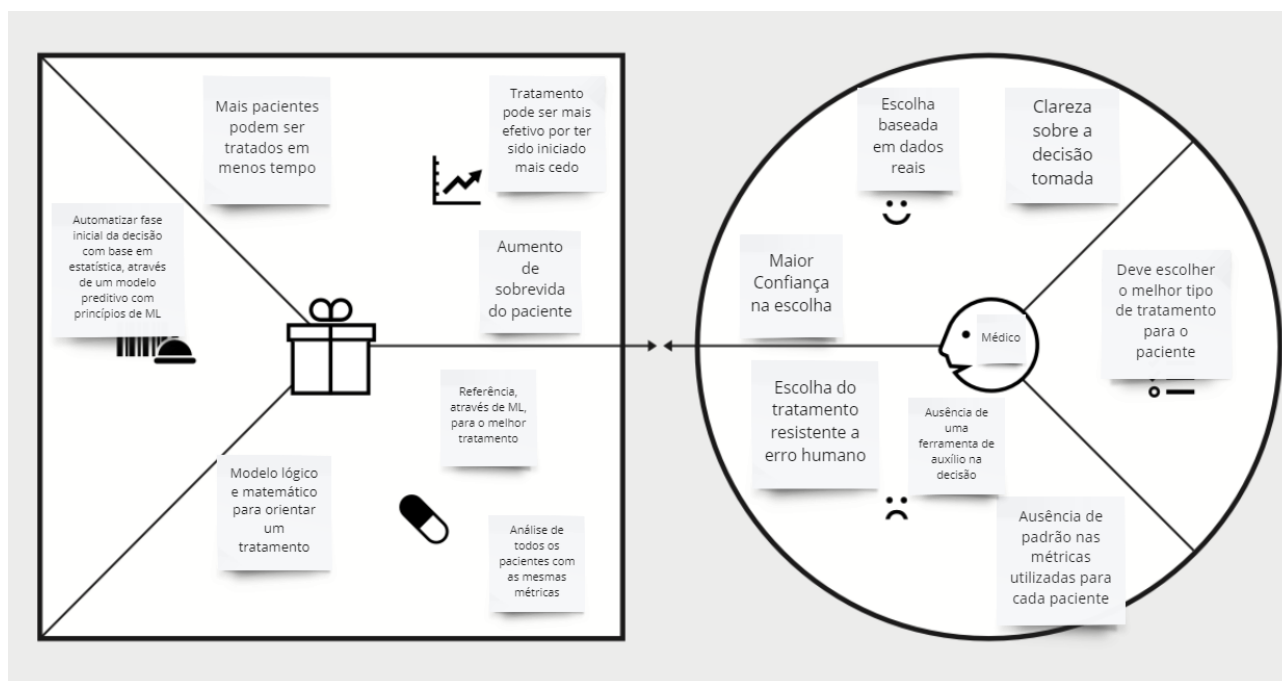


Figura 3 - Value Proposition Canvas

Fonte: Elaboração própria

4.1.5. Matriz de Risco

A importância da matriz de risco para o nosso projeto ajuda a identificar, avaliar e priorizar os riscos potenciais. Isso nos permite ter uma visão clara e objetiva dos desafios e ameaças enfrentados, bem como das ações a serem tomadas para minimizar seu impacto. Além disso, a matriz de risco também fornece uma base para a monitorização contínua dos riscos e para a atualização dos planos de mitigação, garantindo assim a adaptação a mudanças no ambiente do projeto. Em resumo, a tabela de matriz de risco é fundamental para garantir o sucesso do projeto, tomando medidas preventivas para mitigar ameaças e maximizar oportunidades.

Figura 4 - Matriz de Risco.

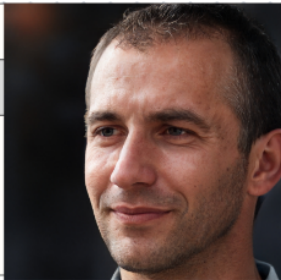
Probabilidade	Ameaças	Oportunidades	Probabilidade
---------------	---------	---------------	---------------


90%											90%
70%				F			I				70%
50%				G							50%
30%		A	D		H						30%
10%			B	C	E						10%
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo	
Impacto											

	NOME	CATEGORIA	PROBABILIDADE	IMPACTO
A	Descompromisso com o horário de desenvolvimento do projeto	Desenvolvimento	30%	BAIXO
B	Desentendimento entre os membros da equipe	Comunicação	10%	MODERADO
C	Médicos não fazerem a utilização do modelo preditivo	Desenvolvimento	10%	ALTO
D	Falta de dados complementares	Comunicação	30%	MODERADO
E	Falta de dados essenciais	Comunicação	10%	MUITO ALTO
F	Efeitos colaterais do tratamento	Desenvolvimento	70%	ALTO
G	Abandono do tratamento pelo paciente	Desenvolvimento	50%	ALTO
H	Os tratamentos não fazerem efeitos	Desenvolvimento	30%	MUITO ALTO
I	Ajudar na escolha do melhor tratamento para a paciente diagnosticada com câncer de mama através do modelo preditivo	Desenvolvimento	70%	ALTO

Fonte: Elaboração própria.

4.1.7. Personas

NOME:	Marcos Fernandes Neto	
INFORMAÇÕES PESSOAIS:		
<ul style="list-style-type: none">• Marcos possui 52 anos.• Formado em medicina.• Trabalha como médico há 22 anos.• É casado e possui 2 filhos.• Nasceu e mora em São Paulo.		
DORES:	<ul style="list-style-type: none">• Não sabe para qual tipo de tratamento encaminhar seus pacientes.• Às vezes seus pacientes não voltam para dar continuidade ao tratamento.• Muitos pacientes só procuram atendimento médico quando o câncer já está em um estágio avançado.	
OBJETIVOS/NECESSIDADES:		
<ul style="list-style-type: none">• Diminuir a quantidade de casos que necessitam tratamentos mais severos e invasivos.• Um modelo preditivo para que ele saiba para qual tratamento encaminhar o paciente.• Deseja saber qual é o tratamento que trará mais resultados aos pacientes.		

NOME:	Renata Gonçalves Dias	
INFORMAÇÕES PESSOAIS:		
<ul style="list-style-type: none"> • Renata possui 41 anos. • Está fazendo mestrado em medicina. • Trabalha como pesquisadora. • É casada. • Nasceu na Bahia e mora em São Paulo. 		
DORES:	<ul style="list-style-type: none"> • Não sabe porque alguns pacientes respondem melhor ao tratamento neo enquanto outros respondem melhor ao tratamento adjuvante. • Os dados frequentemente possuem informações vazias ou são insuficientes. • Desconhece o que influencia na taxa de sucesso dos tratamentos. 	
OBJETIVOS/NECESSIDADES:		
<ul style="list-style-type: none"> • Saber quais fatores fazem o paciente responder melhor a determinado tratamento. • Explorar meios que viabilizem uma diminuição dos efeitos colaterais causados pelos tratamentos. • Deseja descobrir se a alteração na sequência dos processos causa uma diminuição no tempo total de tratamento do paciente. 		

Ambas personas utilizarão o modelo e serão afetadas por ele no sentido laboral, de maneira em que os resultados desse modelo podem afetar diretamente em seus trabalhos, já que pode mudar completamente a visão desses profissionais (médico e pesquisadora) a respeito do tratamento do câncer de mama.

4.1.6. Jornadas do Usuário

O Mapa de Jornada do Usuário é uma ferramenta que ajuda a entender e acompanhar as fases pelas quais um usuário passa ao interagir com determinado modelo. Nesse caso, os usuários são médicos interagindo com a plataforma que realiza análises e previsões sobre qual tratamento é ideal para cada paciente.

A primeira fase é o Conhecimento da Plataforma. Nesta etapa, o médico entra em contato com o modelo pela primeira vez e precisa compreender o objetivo e o funcionamento da plataforma para que possa utilizá-la da melhor forma possível. É importante que ele saiba manusear a plataforma e entenda suas funcionalidades.

Na segunda fase, a Entrada de Dados, o médico precisa inserir informações sobre o paciente, como dados clínicos, exames, histórico médico, etc., para que o modelo possa fazer a análise e gerar uma predição. É importante que o médico preste atenção aos dados inseridos para garantir a precisão da análise.

Na terceira fase, a Saída de Informações, o médico tem acesso ao resultado gerado pelo modelo, ou seja, à predição. É importante que ele entenda o porquê da predição e que possa interpretar corretamente as informações geradas.

Por fim, a quarta e última fase é o Final do Tratamento, momento em que o médico informa a conclusão do procedimento, fornecendo dados para o modelo aprender e aumentar sua acurácia. É importante que ele preste atenção aos resultados finais para que possa contribuir para o desenvolvimento da plataforma.

Em resumo, o Mapa de Jornada do Usuário é uma ferramenta valiosa para entender e acompanhar o processo de interação do médico; o principal usuário, com a plataforma, garantindo que ele possa utilizá-la da maneira mais eficiente e eficaz.

Figura 5 - Mapa de Jornada de Usuário.

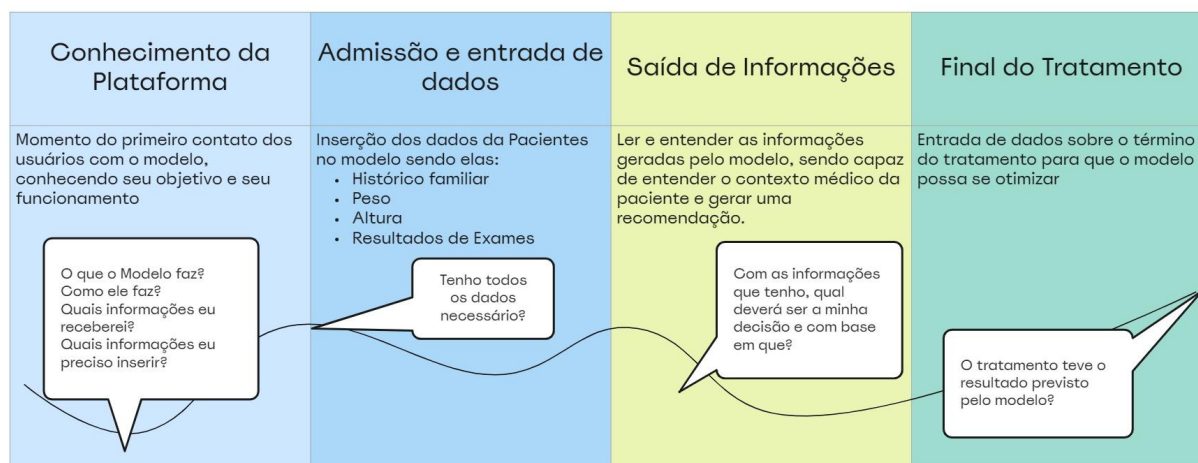


Marcos Fernandes Neto - Médico

Cenário: Com o diagnóstico do paciente, ele precisa saber qual o melhor tratamento para a mesma

Expectativas

Ele espera ser capaz de entender informações que sejam conclusivas para que ele possa indicar o tratamento mais apropriado



Oportunidades

- Ser uma plataforma de fácil e rápido entendimento para ágil aprendizado pelos usuários
- Ter um relatório detalhado para fundamentar a lógica do modelo e facilitar o entendimento do médico sobre o por quê do resultado

Responsabilidades

Garantir que as informações sejam claras, facilmente inseridas e entendidas pelo corpo médico

miro

Fonte: Elaboração própria.

4.1.7 Política de privacidade para o projeto de acordo com a LGPD

A Pink Solution é um grupo focado na análise de dados médicos - providos do Instituto do Câncer do Estado de São Paulo (ICESP) - para a composição de um modelo preditivo para auxiliar os médicos por meio de um prognóstico de qual dos tratamentos para câncer de mama, entre adjuvante e neoadjuvante, deveria ser recomendado ao paciente.

Nós, da Pink Solution, somos comprometidos em proteger a privacidade e segurança dos dados clínicos de nossos pacientes. Esta política de privacidade descreve como coletamos, usamos, armazenamos e compartilhamos informações pessoais sensíveis, incluindo dados clínicos, na operação deste projeto de recomendação de tratamentos médicos.

Coleta de Dados Clínicos:

Recebemos dados clínicos dos pacientes através da base de dados do Instituto do Câncer do Estado de São Paulo (Icesp). Esses dados são coletados com o consentimento dos pacientes e são usados exclusivamente para fins médicos. As informações coletadas consistem em: idade, sexo, raça declarada, peso, altura, IMC, escolaridade, informações de estado: vivo ou óbito, informações pessoais (gravidez, menstruação, utilização de métodos contraceptivos e uso de drogas), histórico familiar de câncer e estado clínico do paciente.

Uso de Dados Clínicos:

Os dados clínicos coletados são usados exclusivamente para fornecer recomendações de tratamento precisas e personalizadas aos médicos, ajudando-os a tomar decisões que forneçam a maior probabilidade de sucesso no tratamento do paciente.

Armazenamento de Dados Clínicos:

Os dados Clínicos são armazenados em servidores seguros, protegidos por medidas de segurança físicas e digitais de alta qualidade. O acesso a esses dados é limitado a pessoal autorizado com uma necessidade legítima de conhecer essas informações, como médicos e pesquisadores. A Lei 13.787/18 disciplina a digitalização e a utilização de sistemas informatizados para a guarda, o armazenamento e o manuseio de prontuários de pacientes e o tempo de guarda dos prontuários médicos corresponde a 20 anos (MORSCH, 2022).

Compartilhamento de Dados Clínicos:

Os dados Clínicos não serão compartilhados com terceiros, exceto quando exigido por lei ou quando houver uma necessidade médica aparente. Em tais casos, o compartilhamento será feito somente após o devido processo legal e com o consentimento dos pacientes.

Segurança de Dados Clínicos:

Tomamos medidas rigorosas para garantir a segurança e a privacidade dos dados clínicos. Isso inclui a implementação de medidas de segurança físicas e digitais, como criptografia de dados, autenticação de usuário e backup frequente.

Direitos dos Pacientes:

Os pacientes têm o direito de acessar, corrigir e excluir seus dados clínicos a qualquer momento. Para exercer esses direitos, os pacientes devem entrar em contato através dos meios fornecidos na página de contato do ICESP.

4.2. Compreensão dos Dados

1. Exploração de dados:

a) Cite quais são as colunas numéricas e categóricas.

Colunas numéricas contêm valores numéricos, ou seja, valores que representam números. Colunas numéricas podem representar, por exemplo, idade, peso, altura, temperatura, etc.. Esses valores são contínuos quando contemplam uma gama de valores possíveis, ou discretos, quando contemplam apenas valores separados e distintos.

Colunas categóricas, por outro lado, contêm valores que representam categorias: sexo, cor dos olhos, estado civil, etc.. Esses valores são representados por strings ou códigos que indicam a categoria a que pertencem.

Uma forma simples de identificar se uma coluna é numérica ou categórica é observar os valores presentes na coluna. Se a maioria dos valores for números, a coluna é provavelmente numérica. Se a maioria dos valores for palavras ou frases, a coluna é provavelmente categórica.

Outra forma de identificar é utilizando funções de programação que permitam analisar os dados, como a função *describe()* no Python, que retorna um resumo estatístico de colunas numéricas, ou a função *unique()* que retorna os valores únicos presentes em colunas categóricas.

Para reconhecer se são colunas numéricas ou colunas categóricas nós utilizamos o código abaixo para otimizarmos o tempo e conseguirmos verificar sem precisar olhar precisamente os

dados. Explicando o código, caso a coluna seja igual a 'float64' ou 'int64' considera o tipo como numérico, e se a coluna for igual a 'object' o tipo será categórico.

RECONHECIMENTO DE COLUNAS NUMÉRICAS X COLUNAS CATEGÓRICAS

[+ Code](#)[+ Markdown](#)

```
cont = 0
listacolcat = []
for coluna in tes2.dtypes:
    if coluna == 'float64' or coluna == 'int64':
        tipo = "Coluna Numérica"

    elif coluna == object:
        tipo = "Coluna Categórica"
        listacolcat.append(tes2.columns[cont])

    print(f'{tes2.columns[cont]} é {tipo}\n=====')
    cont+=1
```

Por fim, a próxima imagem retrata algumas respostas de como seria o output do código acima.

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

Record ID é Coluna Numérica

=====

Idade do paciente ao primeiro diagnóstico é Coluna Numérica

=====

Última informação do paciente é Coluna Categórica

=====

Lista de colunas numéricas

- Record ID
- Idade do paciente ao primeiro diagnóstico
- Data da última informação sobre o paciente
- Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci]
- Quantas vezes ficou grávida?
- Idade na primeira gestação
- Por quanto tempo amamentou?
- Data da cirurgia
- Data de início do tratamento quimioterapia
- Data do início Hormonoterapia adjuvante
- Data do diagnóstico
- Data de início da Radioterapia
- Grau histológico
- Subtipo tumoral
- Receptor de progesterona (quantificação %)
- Receptor de Estrogênio (quantificação %)
- Índice H (Receptor de progesterona)
- Data do tratamento
- IMC
- Data de Recidiva
- Ki67 (%)
- Data:
- Ano do diagnóstico
- Peso
- Altura (em centímetros)
- Data da primeira consulta institucional [dt_pci]
- Código da Morfologia de acordo com o CID-O

Lista de Colunas Categóricas

- Já ficou grávida?
- Ultima_informacao_paciente
- Amamentou na primeira gestação?
- Atividade Física
- Regime de Tratamento
- Tipo de terapia anti-HER2 neoadjuvante
- Radioterapia
- Esquema de hormonioterapia
- Diagnostico primario (tipo histológico)
- Receptor de estrogênio
- Receptor de progesterona
- Ki67 (>14%)
- HER2 por IHC
- HER2 por FISH
- Código da Topografia (CID-O)
- Estadio Clínico
- Grupo de Estadio Clínico
- Classificação TNM Clínico - T
- Classificação TNM Clínico - N
- Classificação TNM Clínico - M
- Combinação dos Tratamentos Realizados no Hospital
- Lateralidade do tumor
- Local de Recidiva a\xa0 distancia/ metastase #1 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #2 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #3 - CID-O - Topografia
- Local de Recidiva a\xa0 distancia/ metastase #4 - CID-O - Topografia
- Com recidiva à distância
- Com recidiva regional
- Com recidiva local

b) Estatística descritiva das colunas.

Usamos a função `describe()` o qual é um método do objeto Data Frame do Pandas, que retorna um conjunto de estatísticas descritivas para as colunas numéricas do Data Frame. Essas estatísticas incluem a contagem de valores não nulos, a média (Mean), o desvio padrão (Std), o valor mínimo (Min) e máximo (Max), o primeiro quartil (25%), a mediana (50%) e o terceiro quartil (75%).

Para as colunas que contêm dados não numéricos, a função `describe()` não é aplicável, pois essas estatísticas não têm significado para esses tipos de dados. Nesses casos, é possível usar outros métodos do Pandas, como `value_counts()`, `unique()`, `nunique()` ou `groupby()`, dependendo do que se deseja analisar.

Para fazer a estatística descritiva, nós selecionamos apenas algumas colunas numéricas e categóricas para mostrarmos de exemplo no documento.

Numéricas

- "Idade do paciente ao primeiro diagnóstico";
- "Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos".

ESTATÍSTICA DESCRITIVA DAS COLUNAS NUMÉRICAS			
<pre>tes2.describe()</pre>			
	Record ID	Idade do paciente ao primeiro diagnóstico \	
count	3726		3726
mean	49219		54
std	20989		13
min	302		22
25%	30608		45
50%	54701		54
75%	67576		63
max	82240		89

```
Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos    [dt_pci]
count                                          3726
mean                                          1501
std                                           842
min                                           25
25%                                          1008
50%                                          1301
75%                                          1837
max                                          4503
```

Catégoricas

Para fazer a estatística descritiva das colunas catégoricas utilizamos o método `.value_counts()` o qual é uma função em Python que pode ser aplicada a uma série de dados. Ele retorna uma contagem de valores únicos na série e a frequência de cada valor. A saída do `.value_counts()` é uma lista com índices correspondentes aos valores únicos encontrados na série de entrada e valores correspondentes à contagem de cada valor único na série de entrada. Essa contagem é classificada em ordem decrescente de frequência. Esse método serve para entender a distribuição de valores em uma série de dados e pode ser usado para análises exploratórias de dados.

Colunas usadas para a análise descritiva:

- “Última informação do paciente”;
- “Já ficou grávida?”;
- “Regime de Tratamento”

ESTATÍSTICA DESCRITIVA DAS COLUNAS CATEGÓRICAS

```
for coluna in listacolcat:
    print(f'Coluna:{coluna}\n\n{tes2[coluna].value_counts()}\n\n')
```

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

Coluna:Última informação do paciente

Vivo, SOE	2536
Óbito por câncer	910
Vivo, com câncer	218
Óbito por outras causas, SOE	62

Name: Última informação do paciente, dtype: int64

Coluna:Já ficou grávida?

Não Informado Gravida	2787
Sim	928
Não	11

Name: Já ficou grávida?, dtype: int64

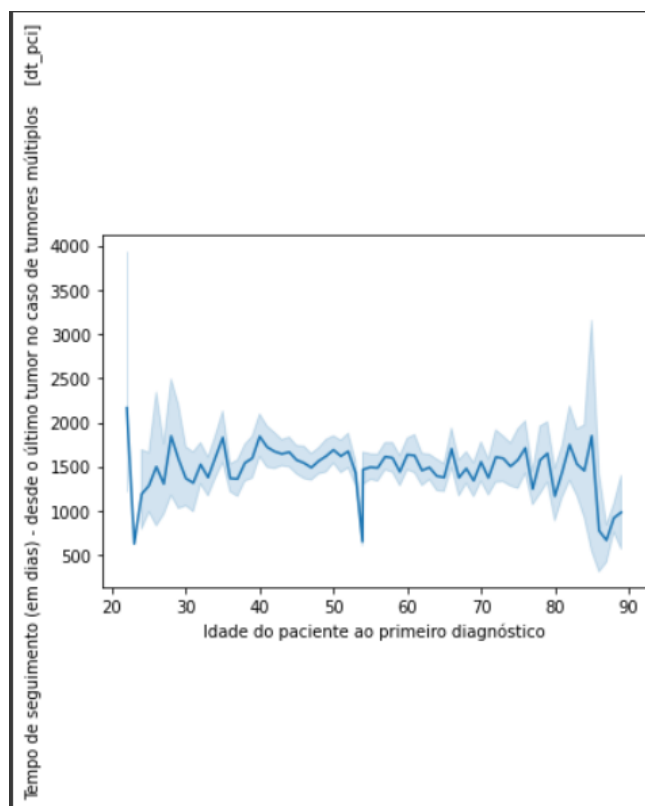
Coluna:Regime de Tratamento

Terapia Adjuvante	1275
Não Informado Tratamento	1195
Terapia Neoadjuvante	1176
Paliativo	55
...	

Name: Combinação dos Tratamentos Realizados no Hospital, dtype: int64

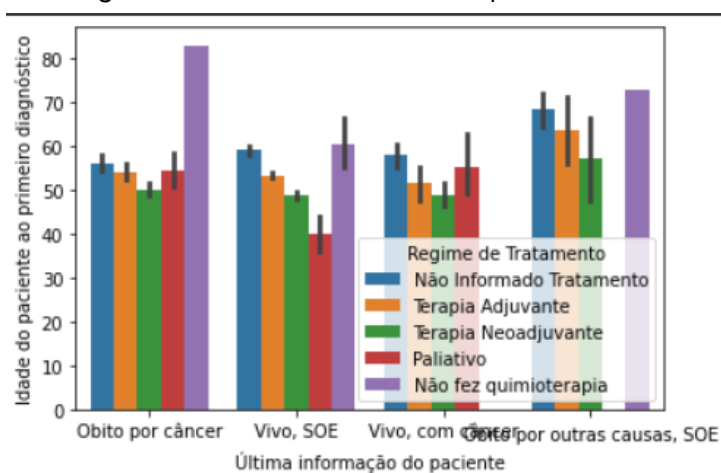
Gráficos relacionais entre variáveis escolhidas pelo grupo

Neste primeiro gráfico relacionamos a idade do paciente com o tempo desde o último tumor.



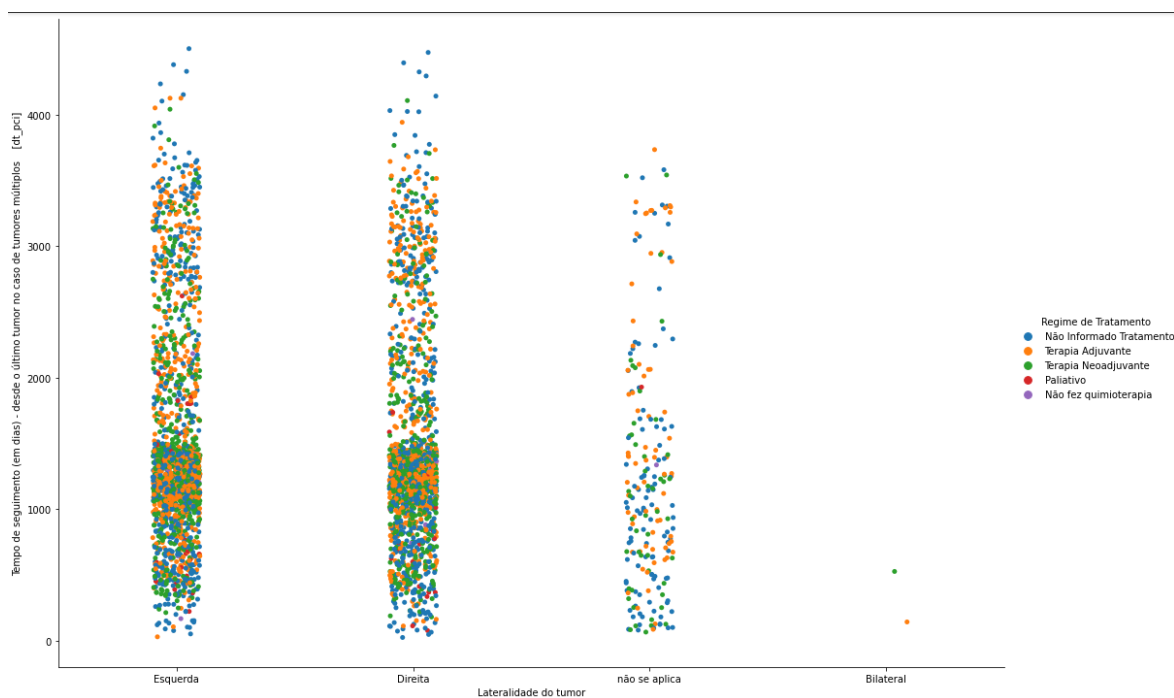
Fonte: Elaboração própria

No segundo gráfico relacionamos a situação do paciente com sua idade e explicitamos o regime de tratamento utilizado por cada um.



Fonte: Elaboração própria

Por último, este gráfico tenta enxergar a relação entre a lateralidade do tumor e o tempo desde o último tumor, novamente tentando explicitar qual foi o regime de tratamento utilizado.



Fonte: Elaboração própria

2. Pré-processamento dos dados:

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

Durante a análise do tratamento de dados, foram excluídos os IMC's menores que 5 e acima de 50, pois, sendo altamente improvável que um paciente alcance tal nível, a acurácia do modelo seria negativamente afetada.

Por meio da mesma lógica, na coluna "Tempo desde o ultimo tumor", retirou-se os que marcavam abaixo de 20 dias por virtude do fato de que 20 dias é um intervalo de tempo demasiadamente baixo.

Também consideramos a exclusão de pacientes com idades acima de 90 e abaixo de 18 foram desconsideradas porque eram poucos registros e/ou não-práticos - algumas chegavam a

marcar mais de 100 anos e outras menos de 1. Para a construção do modelo foram considerados apenas adultos.

```
[ ] #SETANDO O NUMERO DE CASAS DECIMAIS
pd.set_option('display.precision',1)

menores = df_peso[(df_peso.IMC > 5)]
maiores = menores[(menores.IMC < 50)]

#Agrupando por ID e colocando a média
df_peso = maiores[(maiores.IMC != np.inf)].g

#pegar só a ultima ocorrencia
```

Vale ressaltar que no pré processamento dos dados, devido ao modelo de gravação dos dados, uma mesma pessoa (Record ID) teve diversas linhas registrando suas variações de peso e altura. Para agruparmos os Record ID em uma única linha, mantivemos a última linha registrada de cada índice.

RETIRANDO LINHAS COM MAIS DE UMA OCORRÊNCIA, MANTENDO O ÚLTIMO REGISTRO

```
[ ]

df_hist = df_hist.drop_duplicates(subset=['R
df_tumor = df_tumor.drop_duplicates(subset=[

print(df_peso['Record ID'].nunique())
print(df_hist['Record ID'].nunique())
print(df_tumor['Record ID'].nunique())
print(df_demo['Record ID'].nunique())

print('-----')

print(df_peso['Record ID'].value_counts().su
print(df_hist['Record ID'].value_counts().su
print(df_tumor['Record ID'].value_counts().s
print(df_demo['Record ID'].value_counts().su
```

3. Hipóteses:

Por meio da estratificação dos pacientes conforme a jornada de tratamento e da identificação de qual estágio ele se encontra, foi possível criar três hipóteses.

a) Levantamento das três hipóteses com justificativa.

(As hipóteses foram formuladas conforme a análise de dados com um escopo menor. Por virtude do fato de que o modelo preditivo criado foca em apenas dois tratamentos, ("Adjuvante" e "Neoadjuvante") as outras opções ("paliativo" e "não fez quimioterapia") foram descartadas; foram analisadas principalmente os padrões dos estádios clínicos mais avançados (i.e. IIIA, IIIB, IIIC e IV), pois foram os resultados que deram mais divergência entre os dois tipos de tratamento. Na coluna "última informação do paciente", foram consideradas as informações mais objetivas: "morte por câncer" e "vivo SOE". Não foram utilizadas as informações de "óbito por outras causas", pois impacta negativamente na análise, e nem "vivo com câncer", pois existe uma série de outros fatores que podem influenciar na presença do câncer, mesmo após o tratamento).

O objetivo principal é, no mínimo, prolongar a vida do paciente, e, no máximo curá-lo do câncer. Por isso, a principal métrica de sucesso utilizada foi o tempo de sobrevida do paciente.

As variáveis utilizadas são: Regime de Tratamento (Adjuvante e Neoadjuvante), última informação do paciente (as informações de "óbito por câncer" e "vivo SOE"), faixa etária*, estágio do câncer (IIIA, IIIB, IIIC e IV), período de tratamento**, contagem de Record ID e metástase***.

Primeira Hipótese:

Para as pessoas do grupo IIIA, o tratamento mais adequado seria o Adjuvante (independente da faixa etária).

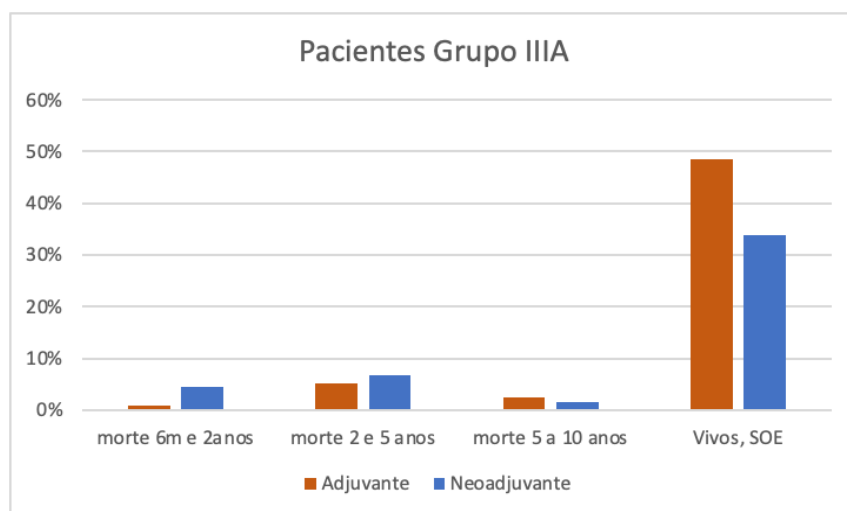
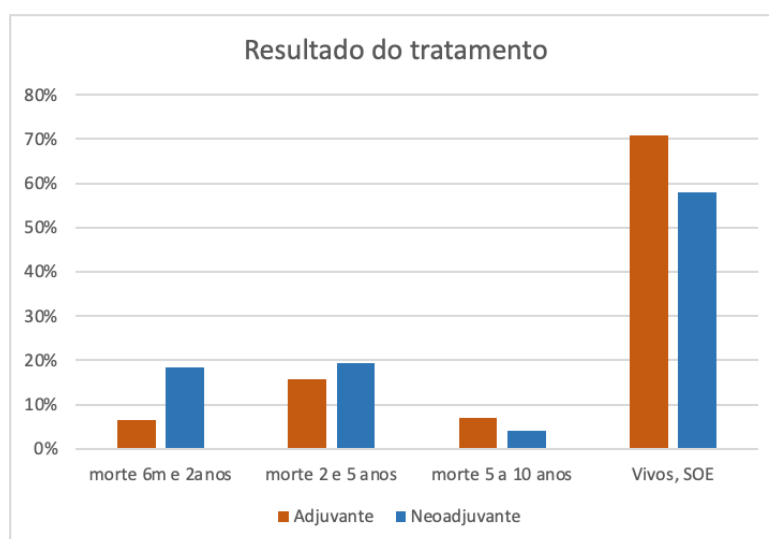


Figura 6 - Pacientes Grupo IIIA (% de mortalidade)¹

Fonte: Elaboração própria.

Segunda Hipótese:

No período entre 6 meses e 2 anos, o falecimento de pacientes com até 60 anos é muito superior na terapia Neoadjuvante. Na terapia Adjuvante há uma porcentagem de falecimento de 4%, enquanto na terapia Neoadjuvante há uma porcentagem de 13%.



¹ Para mais informações (segunda tabela da página "PIVOT ANALISE"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=120895560>

Figura 7 - Resultado do Tratamento (% de mortalidade)²

Fonte: Elaboração própria.

Terceira Hipótese:

A diferença entre a porcentagem de ter metástase em pacientes de 40 a 60 anos é maior na terapia Neoadjuvante, principalmente no período de 6 meses a 2 anos (Adjuvante - 0,76% ; Neoadjuvante - 2,53%).

Morte metástase	Total	40 < ID <=60
Adjuvante	4,29%	0,76%
Neoadjuvante	5,52%	2,53%

Figura 8 - Mapa de Jornada de Usuário.³

Fonte: Elaboração própria.

*Faixa etária: uma coluna criada a partir da informação das idades de todos os pacientes que foram subdivididos em três grupos: menos de 40 anos, entre 40 e 60 anos e mais de 60 anos.

**Período de tratamento: foi calculado o período do tratamento, primeiramente em dias, por meio da subtração da "Data da Última informação do paciente" pela "Data do tratamento". Com uma coluna com o tempo do tratamento contado em dias, foi criada outra coluna que a subdividiu em cinco setores: menos de 180 dias (menos de 6 meses), entre 180 dias e 2 anos, entre 2 a 5 anos, entre 5 a 10 anos e mais de 10 anos.

***Metástase: foi considerada a primeira coluna de presença de metástase: "Metastase ao DIAGNOSTICO - CID-O #1", para criar uma coluna de "Metastase ou não", definida pela presença de metástase ou não, independentemente de onde o câncer foi identificado no corpo.

²Para mais informações (segunda tabela da página "PIVOT ANALISE"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=120895560>

³ Para mais informações (tabelas da página "PIVOT Análise 2"):

<https://docs.google.com/spreadsheets/d/1-5LJexvUS2a7eD-svDnMGLIHfjtuZLBu/edit#gid=909078358>

4.3. Preparação dos Dados e Modelagem

4.3.1. Modelagem para o problema

Como mencionado na Introdução, o projeto visa desenvolver inteligência artificial que, com relativo alto grau de acurácia, prevê o tratamento probabilisticamente ideal para cada paciente de câncer de mama - uma probabilidade que é adquirida por meio da meticulosa análise de múltiplos dados subproduto de detalhados relatórios sobre milhares de pacientes. Abaixo, foram detalhados os dados que possuem maior influência na escolha do tratamento.

- **Record ID:** utilizado para o tratamento dos dados - identificação de cada paciente e suas respectivas condições.
- **Idade do paciente ao primeiro diagnóstico:** é um fator importante para metrificar qual a faixa etária dos pacientes tratados. Pacientes idosos, por exemplo, podem ser inaptos ao tratamento, mas, por outro lado, pacientes mais novos podem apresentar cânceres com deterioração acelerada.
- **Última informação do paciente:** utilizada para a identificação do estado* do paciente, é uma informação fundamental para a análise e verificação de sucesso ou fracasso do tratamento.

*tal estado pode ser "Vivo SOE"; "Vivo com câncer"; "Óbito por câncer"; "Óbito POC".

- **Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci]:** utilizado para verificar quão recente é o tumor e qual impacto teve no tratamento. Tumores mais recentes apresentam maiores chances de serem curados.
- **Já ficou grávida?:** utilizada para verificar se a gravidez influencia na escolha do tratamento, pois pacientes previamente grávidas provavelmente amamentaram, e amamentação evidencia o desenvolvimento completo da mama.
- **Regime de Tratamento:** define o tipo de tratamento escolhido (Não fez quimioterapia, Paliativo, Terapia Adjuvante e Terapia Neoadjuvante).
- **Classificação* TNM Clínico - M:** indica a existência da presença de metástase em outros órgãos, um fator que define a complexidade do câncer. Cânceres mais complexos demandam tratamentos mais complexos.

*As pessoas que apresentaram metástase antes do diagnóstico são indicadas para o tratamento paliativo. Portanto, nos tratamentos Adjuvante ou Neoadjuvante, significa o surgimento de metástase durante o tratamento.

- **Classificação TNM Clínico - N:** descreve se existe disseminação da doença para os linfonodos regionais; se sim, significa que o câncer começou a atacar o sistema imunológico.

- **Classificação TNM Clínico - T:** indica o tamanho do tumor primário, que influencia no estágio clínico do câncer, que pode influenciar na escolha do tratamento. (o estágio pode ser ou inicial ou avançado)
- **Lateralidade do tumor:** verifica se o local onde o tumor está localizado interfere na escolha do tratamento.
- **Com recidiva à distância, Com recidiva regional, Com recidiva local:** como mencionado na sprint 2 pela líder executiva Luciana, a presença de recidiva é um fator crítico para definir o fracasso do tratamento.
- **Estadio Clínico:** utilizado para definir e diferenciar os estádios do câncer (I, IA, IB, II, IIA, IIB, IIIA, IIIB, IIIC, IV, IVB) de cada paciente, é importante para saber qual as diferenças nas recomendações de tratamento dependendo do estágio - a recomendação de tratamentos pode variar se o estágio do câncer for avançado ou inicial.
- **Combinação dos Tratamentos Realizados no Hospital:** utilizado para verificar se outros tratamentos - radioterapia, hormonioterapia ou outras combinações - interferem na escolha do tratamento.

4.3.2. Métricas relacionadas ao modelo

Dentre todas as métricas apresentadas, optou-se por utilizar 4: acurácia, precisão, recall e f1-score.

A acurácia mede a quantidade de acertos nas previsões em relação a todas as previsões, ou seja, uma alta acurácia indica que há poucos casos de predições erradas, sejam elas falso positivo ou falso negativo. Tal métrica indica o quão assertivo, de modo geral, é o modelo.

Diferentemente da acurácia, a precisão indica o quão assertivo o modelo é em relação às suas predições. Isto é, indica a porcentagem de previsões corretas dentre todas as previsões "positivas". A métrica se aplica ao projeto pois retorna o grau de confiabilidade da predição de sucesso para determinado tratamento.

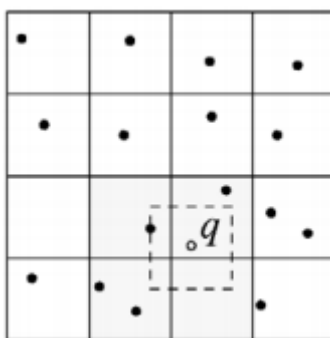
Por outro lado, tem-se também o Recall ou Sensibilidade que retorna a relação das previsões verdadeiras em relação a todos os casos "positivos", ou seja, retorna no modelo o quanto ele deixou de prever corretamente. Dessa forma, um recall indica que o modelo prevê um falso negativo poucas vezes - isto é, deixa de prever um caso positivo.

Por fim, ressalta-se o f1 score, que busca retornar um equilíbrio entre a precisão e o recall, através de uma média harmônica das duas métricas. Dessa forma, o f1 score funciona como uma métrica que indica o quão bem o modelo funciona de modo geral, mas de forma ainda mais complexa.

4.3.3. Modelos Comparados

- **KNN:** KNN (K-Nearest Neighbors, em tradução literal “K-vizinhos mais próximos”) é um algoritmo para reconhecimento de padrões utilizando de método não-paramétrico que, basicamente, classifica a base de dados em dados para treino e em dados para testes. A distância entre os pontos de treino e os pontos de testes é avaliada, e o ponto com a menor distância é classificado como o nearest neighbor. O algoritmo KNN prevê o resultado com base na maioria, como demonstra a imagem abaixo.

A principal desvantagem do KNN, se comparado a outros modelos, resume-se na possibilidade de que o período de processamento seja desconfortavelmente longo, posto que a quantidade de tempo de processamento aumenta de acordo com o tamanho da base de dados.



Fonte: MIT OpenCourseWare

- **Naïve Bayes:** é um algoritmo supervisionado de classificação. O Naïve Bayes pertence à ordem de algoritmos de generative learning, ou seja, aqueles que modelam a distribuição de inputs de uma determinada classe ou categoria.

Também rotulado como classificador probabilístico, o Naïve Bayes é estruturado sobre o Teorema de Bayes. Essencialmente, o Teorema de Bayes permite que probabilidades condicionais sejam invertidas. (Probabilidades condicionais representam a probabilidade da ocorrência de determinado evento dada a ocorrência prévia de outro evento.) O Teorema de Bayes é representado pela seguinte fórmula:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

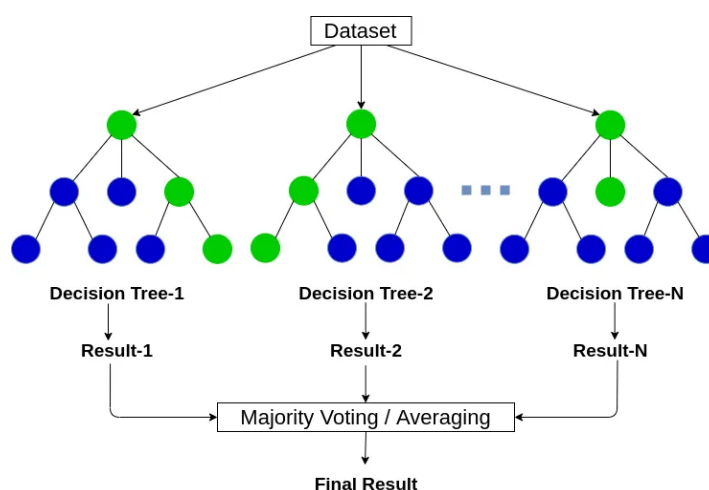
Fonte: Autoria própria

O Teorema de Bayes destaca-se por virtude do fato de que utiliza eventos sequenciais de maneira que informações adquiridas posteriormente conseguem impactar a

probabilidade inicial. Tais probabilidades são indicadas como anterior ou posterior. A probabilidade anterior é a probabilidade inicial de um evento antes de ser contextualizado sob uma determinada condição, ou a probabilidade marginal. A probabilidade posterior é a probabilidade de um evento após observar certo dado.

A principal falha do Naïve Bayes reside no fato de que é incapaz de aprender quais características são mais importantes para diferenciar entre classes.

- **Random Forest:** modelo que, como indica o nome, consiste em um conjunto de árvores de decisões: um número n de árvores com um número x de ramificações, sendo os valores de n e de x decididos pelos desenvolvedores. Tanto no Neoadjuvante quanto no Adjuvante, após escolhidas um máximo de 7 ramificações, o modelo contou as respostas das árvores e retornou aquela que mais obteve respostas segundo as árvores.



[Fonte:](#)

4.3.4. Apresentação do primeiro modelo candidato e Discussão sobre os resultados desse modelo

Por virtude dos resultados obtidos em cada um dos métodos, e por consequência de sua lógica de funcionamento, o modelo escolhido foi o Random Forest. Inicialmente, utilizou-se os métodos K-Nearest Neighbors (KNN), Naïve Bayes e Random Forest, enquanto foram feitos diversos testes de hipóteses.

A cada alteração, foram aplicados os três métodos, e comparados os resultados. Ao final, definiu-se um modelo que teve seu target criado de acordo com uma lógica baseada na coluna “Última informação do paciente”, que tinha como domínio “Vivo, SOE”, “Vivo, com câncer”, “Óbito

por câncer”, “Óbito por outras causas, SOE”. A coluna target, ou seja, coluna que o modelo tenta prever, recebe apenas 0 ou 1: insucesso e sucesso, respectivamente.

Com isso em mente, foram descartadas as linhas nas quais a célula era “Óbito por outras causas, SOE”, pois, a princípio, não é possível inferir se o tratamento foi um sucesso ou não. Então, caso a célula seja “Vivo, SOE” é considerado sucesso; caso a célula seja “Vivo, com câncer” conclui-se que se o paciente estiver no processo de tratamento (sem recidiva), consideramos sucesso; mas, se houver recidiva, considera-se insucesso. Por fim, se a célula for “Óbito por câncer” também entende-se insucesso.

Dessa forma, o modelo é responsável por prever Sucesso ou Insucesso com base nas informações de cada paciente. Em conclusão, utilizou-se dois modelos: um responsável por prever sucesso ou insucesso de um input para o tratamento Adjuvante, e outro para prever sucesso ou insucesso para o tratamento Neoadjuvante. É importante mencionar que o modelo é idêntico, mas cada um deles foi treinado com as pacientes que tiveram o respectivo tratamento, e por isso, cada um é especializado em um tratamento.

Após os testes, foram obtidos os seguintes resultados em cada método:

KNN Adjuvante

Acc treino: 0.87				
Acc teste: 0.828				
	precision	recall	f1-score	support
0.0	0.53	0.20	0.30	44
1.0	0.85	0.96	0.90	206
accuracy			0.83	250
macro avg	0.69	0.58	0.60	250
weighted avg	0.79	0.83	0.80	250

Observando os resultados acima, temos que o modelo KNN Adjuvante obteve:

Precision: 53%

Recall: 20%

F1-Score: 30%

Accuracy: 83%

Estas são as métricas escolhidas para avaliar os modelos e estarão disponíveis na seção 4.4.1.

KNN Neoadjuvante

Acc treino: 0.7980241492864983				
Acc teste: 0.7017543859649122				
	precision	recall	f1-score	support
0.0	0.57	0.51	0.54	78
1.0	0.76	0.80	0.78	150
accuracy			0.70	228
macro avg	0.67	0.66	0.66	228
weighted avg	0.70	0.70	0.70	228

Observando os resultados acima, temos que o modelo KNN Neoadjuvante obteve:

Precision: 57%

Recall: 51%

F1-Score: 54%

Accuracy: 70%

Random Forest Adjuvante

0.885				
0.848				
	precision	recall	f1-score	support
0.0	0.53	0.20	0.30	44
1.0	0.85	0.96	0.90	206
accuracy			0.83	250
macro avg	0.69	0.58	0.60	250
weighted avg	0.79	0.83	0.80	250

Observando os resultados acima, temos que o modelo Random Forest Adjuvante obteve:

Precision: 83%

Recall: 20%

F1-Score: 30%

Accuracy: 83%

Random Forest Neoadjuvante

```
0.8309549945115258
0.7894736842105263
```

	precision	recall	f1-score	support
0.0	0.57	0.51	0.54	78
1.0	0.76	0.80	0.78	150
accuracy			0.70	228
macro avg	0.67	0.66	0.66	228
weighted avg	0.70	0.70	0.70	228

Observando os resultados acima, temos que o modelo Random Forest Neoadjuvante obteve:

Precision: 57%

Recall: 51%

F1-Score: 54%

Accuracy: 70%

Observações:

Naïve Bayes foi descartado por ter apresentado acurácia abaixo de 40%; especula-se que por consequência das colunas selecionadas.

O KNN e o Random Forest obtiveram acurácia maior para o tratamento adjuvante, porém o Random Forest apresenta acurácia ainda maior, sem caracterizar o overfitting. Por isso, escolheu-se o Random Forest como modelo candidato.

4.4. Comparação de Modelos

4.4.1. Escolha da métrica e justificativa

Utilizando como base, a partir da qualidade do modelo, os fatores de maior impacto para o problema, foi possível concluir que a métrica mais importante é a acurácia. Tal conclusão é justificada por dois fatores: a bilateralidade não-hierárquica da predição final (sucesso ou insucesso) de modo que providencia uma visão geral tanto sobre Falsos Positivos quanto Falsos Negativos; e o alto grau de importância que possui a acurácia para avaliar-se a eficácia do

modelo com relativa facilidade de interpretação (especialmente àqueles envolvidos indiretamente no processo de produção que não possuem proficiência em tecnologia).

4.4.2. Modelos Otimizados

Inicialmente, os modelos foram otimizados utilizando Grid Search e Random Search como algoritmos de otimização para hiperparâmetros. Porém, por conta do tempo de execução excessivamente demorado do Grid Search, foi decidido pelo grupo continuar com a otimização utilizando apenas Random Search.

A ideia que fundamenta o GridSearch é a de criar uma grade (grid) com todas as combinações possíveis de hiperparâmetros, avaliar cada combinação, e, por meio da validação cruzada, encontrar o conjunto ideal de hiperparâmetros.

Similarmente, a ideia que fundamenta o RandomSearch é a de realizar amostragem aleatoriamente (random) a partir da distribuição de hiperparâmetros, avaliar cada combinação, e, por meio da validação cruzada, encontrar o conjunto ideal de hiperparâmetros.

Abaixo, screenshots demonstram o código de cada modelo, seguido de análises sobre seus respectivos funcionamentos. (Para informações sobre o funcionamento dos modelos em-si, vide seção 4.3.3..)

KNN

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV

param_grid_knn={
    'n_neighbors':[5,6,10,15,19],
    'weights':['uniform', 'distance'],
    'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'],
    'leaf_size':[10,20,30,40,50],
    'p':[1,2],
    'metric':['euclidean','manhattan','minkowski','chebyshev','mahalanobis'],
}

knn_grid_1=RandomizedSearchCV(knn,param_grid_knn,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)

knn_grid_1.fit(x_treino_knn_neoadjuvante,y_treino_knn_neoadjuvante.squeeze())
```

Métricas:

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	34
1.0	0.79	0.98	0.88	132
accuracy			0.78	166
macro avg	0.40	0.49	0.44	166
weighted avg	0.63	0.78	0.70	166

Análise:

A metodologia utilizada para a construção do modelo preditivo consiste na separação dos dados entre o Adjuvante e o Neoadjuvante. Por isso, vão ser analisados a precisão para cada um separadamente.

A precisão do modelo **KNN para o tratamento Adjuvante** antes da utilização dos hiperparâmetros:

Acc treino:	0.8785185185185185			
Acc teste:	0.8195266272189349			
	precision	recall	f1-score	support
0.0	0.82	0.81	0.81	164
1.0	0.82	0.83	0.83	174
accuracy			0.82	338
macro avg	0.82	0.82	0.82	338
weighted avg	0.82	0.82	0.82	338

Apesar da precisão do modelo ser relativamente alta, por o modelo ser uma predição para a recomendação de um tratamento de câncer, ele deve ter pelo menos -% de acurácia para ser considerado um bom modelo.

Com a implementação do método de hiperparâmetros de Random Search, os seguintes parâmetros foram analisados: "n_neighbors", "weight", "algorithm", "leaf_size", "p" e "metric". Com isso, foram obtidos os melhores parâmetros para cada fator:

```
knn_random_adjuvante=KNeighborsClassifier(leaf_size=10, metric='manhattan',
n_neighbors=10, p=1,
weights='distance')
```

Esses dados foram colocados no modelo de KNN para melhorar a sua precisão, revocação e F1-score. A imagem a seguir mostra a precisão do modelo após a utilização dos hiperparâmetros:

Acc treino: 0.9911111111111112				
Acc teste: 0.8461538461538461				
	precision	recall	f1-score	support
0.0	0.89	0.78	0.83	164
1.0	0.81	0.91	0.86	174
accuracy			0.85	338
macro avg	0.85	0.84	0.84	338
weighted avg	0.85	0.85	0.85	338

(Para a precisão, revocação e F1-score, o primeiro dado seria - e o segundo dado seria -. Ex.: precision: - = 0.89 e - = 0.81)

A precisão do modelo **KNN para o tratamento Neoadjuvante** antes da utilização dos hiperparâmetros:

Acc treino: 0.8825816485225505				
Acc teste: 0.84472049689441				
	precision	recall	f1-score	support
0.0	0.87	0.83	0.85	169
1.0	0.82	0.86	0.84	153
accuracy			0.84	322
macro avg	0.84	0.85	0.84	322
weighted avg	0.85	0.84	0.84	322

Apesar da precisão do modelo ser relativamente alta, por o modelo ser uma predição para a recomendação de um tratamento de câncer, ele deve ter pelo menos 80% de acurácia para ser considerado um bom modelo.

Com a implementação do método de hiperparâmetros de Random Search, os seguintes parâmetros foram analisados: "n_neighbors", "weight", "algorithm", "leaf_size", "p" e "metric". Com isso, foram obtidos os melhores parâmetros para cada fator:

```
knn_random_neoadjuvante=KNeighborsClassifier(leaf_size=10,
metric='manhattan', n_neighbors=6, p=1,
weights='distance')
```

Esses dados foram colocados no modelo de KNN para melhorar a sua precisão, revocação e F1-score. A imagem a seguir mostra a precisão do modelo após a utilização dos hiperparâmetros:

Acc treino: 0.9930015552099534					
Acc teste: 0.8385093167701864					
	precision	recall	f1-score	support	
0.0	0.87	0.81	0.84	169	
1.0	0.81	0.87	0.84	153	
accuracy			0.84	322	
macro avg	0.84	0.84	0.84	322	
weighted avg	0.84	0.84	0.84	322	

(Para a precisão, revocação e F1-score, o primeiro dado seria 0.0 e o segundo dado seria 1.0. Ex.: precision: 0.0 = 0.87 e 1.0 = 0.81)

Em síntese, a implementação de hiperparâmetros GridSearch e RandomSearch em um modelo KNN tende a alicerçar uma melhoria na performance do modelo ao selecionar o melhor conjunto de hiperparâmetros que produzem a mais alta acurácia. Em mais detalhes, o modelo é melhorado das seguintes maneiras:

- Aumento da Precisão: Os hiperparâmetros GridSearch podem afinar parâmetros como o número de nearest neighbors, a métrica de distância, os pesos, etc. Tais parâmetros

tendem a afetar a precisão do modelo, otimizando o número de vizinhos mais próximos e selecionando a melhor métrica de distância para os dados sendo utilizados.

- Melhor generalização: O ajuste dos hiperparâmetros através do GridSearch pode ajudar a reduzir overfitting ou underfitting. (Overfitting ocorre quando o modelo é excessivamente complexo, apresenta bom desempenho no conjunto de treinamento, mas não no conjunto de testes. Underfitting ocorre quando o modelo é excessivamente simples e não consegue capturar a complexidade dos dados.) Ao selecionar os hiperparâmetros ideais, o modelo passa a generalizar melhor dados recém-adicionados.
- Complexidade Reduzida: O modelo KNN pode ser sensível ao número de nearest neighbors, e à medida que o número de vizinhos aumenta, o modelo torna-se mais complexo. Ajustando os hiperparâmetros via GridSearch, torna-se possível encontrar o equilíbrio ideal entre complexidade e precisão.
- Computação mais rápida: à medida que o número de vizinhos aumenta, KNN torna-se cada vez mais exigente computacionalmente. Ao afinar os hiperparâmetros, reduz-se o tempo de computação, melhorando a performance.

SVC

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import RandomForestClassifier

param_grid_svm={
    'C':[1,2,3,4,5,6], 'kernel':['linear', 'poly', 'rbf', 'sigmoid', 'pré-computado'], 'gamma':['scale', 'auto']}

svm_tt=svm.SVC(C=1.0)

svm_grid_neo=RandomizedSearchCV(svm_tt,param_grid_svm,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)

svm_grid_neo.fit(x_treino_svm_neoadjuvante, y_treino_svm_neoadjuvante)
```

Métricas:

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	40
1.0	0.75	1.00	0.86	122
accuracy			0.75	162
macro avg	0.38	0.50	0.43	162
weighted avg	0.57	0.75	0.65	162

Análise:

A metodologia utilizada para a construção do modelo preditivo consiste na separação dos dados entre o Adjuvante e o Neoadjuvante. Por isso, vão ser analisados a precisão para cada um separadamente.

A precisão do modelo **SVC para o tratamento Adjuvante** antes da utilização dos hiperparâmetros:

Acc treino: 0.8785185185185185					
Acc teste: 0.8195266272189349					
	precision	recall	f1-score	support	
0.0	0.82	0.81	0.81	164	
1.0	0.82	0.83	0.83	174	
accuracy			0.82	338	
macro avg	0.82	0.82	0.82	338	
weighted avg	0.82	0.82	0.82	338	

Apesar da precisão do modelo ser relativamente alta, por o modelo ser uma predição para a recomendação de um tratamento de câncer, ele deve ter pelo menos -% de acurácia para ser considerado um bom modelo.

Com a implementação do método de hiperparâmetros de Random Search, os seguintes parâmetros foram analisados: "n_neighbors", "weight", "algorithm", "leaf_size", "p" e "metric". Com isso, foram obtidos os melhores parâmetros para cada fator:

```
knn_random_adjuvante=KNeighborsClassifier(leaf_size=10, metric='manhattan',
n_neighbors=10, p=1,
weights='distance')
```

Esses dados foram colocados no modelo de SVC para melhorar a sua precisão, revocação e F1-score. A imagem a seguir mostra a precisão do modelo após a utilização dos hiperparâmetros:

```

Acc treino:  0.9911111111111112
Acc teste:   0.8461538461538461
              precision    recall  f1-score   support

         0.0         0.89         0.78         0.83         164
         1.0         0.81         0.91         0.86         174

 accuracy                   0.85         338
 macro avg              0.85         0.84         0.84         338
 weighted avg           0.85         0.85         0.85         338

```

(Para a precisão, revocação e F1-score, o primeiro dado seria - e o segundo dado seria -. Ex.: precision: - = 0.89 e - = 0.81)

A precisão do modelo **SVC para o tratamento Neoadjuvante** antes da utilização dos hiperparâmetros:

```

Acc treino:  0.8825816485225505
Acc teste:   0.84472049689441
              precision    recall  f1-score   support

         0.0         0.87         0.83         0.85         169
         1.0         0.82         0.86         0.84         153

 accuracy                   0.84         322
 macro avg              0.84         0.85         0.84         322
 weighted avg           0.85         0.84         0.84         322

```

Apesar da precisão do modelo ser relativamente alta, por o modelo ser uma predição para a recomendação de um tratamento de câncer, ele deve ter pelo menos -% de acurácia para ser considerado um bom modelo.

Com a implementação do método de hiperparâmetros de Random Search, os seguintes parâmetros foram analisados: "n_neighbors", "weight", "algorithm", "leaf_size", "p" e "metric". Com isso, foram obtidos os melhores parâmetros para cada fator:

```
knn_random_neoadjuvante=KNeighborsClassifier(leaf_size=10,
metric='manhattan', n_neighbors=6, p=1,
weights='distance')
```

Esses dados foram colocados no modelo de KNN para melhorar a sua precisão, revocação e F1-score. A imagem a seguir mostra a precisão do modelo após a utilização dos hiperparâmetros:

Acc treino: 0.9930015552099534				
Acc teste: 0.8385093167701864				
	precision	recall	f1-score	support
0.0	0.87	0.81	0.84	169
1.0	0.81	0.87	0.84	153
accuracy			0.84	322
macro avg	0.84	0.84	0.84	322
weighted avg	0.84	0.84	0.84	322

(Para a precisão, revocação e F1-score, o primeiro dado seria - e o segundo dado seria -. Ex.: precision: - = 0.87 e - = 0.81)

Em síntese, a implementação de hiperparâmetros GridSearchCV e RandomSearch em um modelo SVM tende a alicerçar uma melhoria na performance do modelo ao selecionar o melhor conjunto de hiperparâmetros que produzem a mais alta acurácia. Posto que o SVM tem vários hiperparâmetros que podem ser ajustados, como kernel type, regularization parameter (C), e gamma, são estas as maneiras em que o GridSearchCV melhora o SVM:

- Encontra a combinação ideal de hiperparâmetros: ao performar um search de uma quantidade de hiperparâmetros especificada pelo desenvolvedor, o GridSearchCV, via validação cruzada, avalia cada combinação de hiperparâmetros, e retorna os hiperparâmetros que têm o melhor desempenho nos dados de validação.
- Reduz overfitting: por meio da validação cruzada, o GridSearchCV seleciona os hiperparâmetros que têm o melhor desempenho em dados invisíveis.
- Melhora a precisão: ao encontrar os hiperparâmetros ideais, o GridSearchCV pode melhorar a precisão do modelo SVM em dados de teste.

Random Forest

Hiperparâmetros:

```
from sklearn.model_selection import RandomizedSearchCV

param_grid_rf={
    'n_estimators':[7,10,15,25,40,100,200,300,400,500], 'criterion':['gini',"entropy"],\
    'max_depth':[10,30,50,70,100], 'min_samples_split':[10,20,30,40,50,60],\
    'min_samples_leaf':[2,5,10], 'max_features':[10,30,50]
}
rf_tt=RandomForestClassifier()

rf1_grid=RandomizedSearchCV(rf_tt,param_grid_rf,scoring="accuracy",return_train_score=True,verbose=True,n_jobs=-1)
rf1_grid.fit(x_treino_rf_adjuvante, y_treino_rf_adjuvante)
```

Métricas:

	precision	recall	f1-score	support
0.0	0.33	0.03	0.05	34
1.0	0.80	0.98	0.88	132
accuracy			0.79	166
macro avg	0.57	0.51	0.47	166
weighted avg	0.70	0.79	0.71	166

Análise:

Em síntese, a implementação de hiperparâmetros GridSearch e RandomSearch em um modelo random forest tende a alicerçar uma melhoria na performance do modelo ao selecionar o melhor conjunto de hiperparâmetros que produzem a mais alta acurácia. As maneiras em que o GridSearch melhora o random forest são idênticas àquelas que melhoram o SVC: encontra a combinação ideal de hiperparâmetros, reduz o overfitting e melhora a precisão. Vide a análise prévia para consultá-las em mais detalhes.

4.4.3. Definição do modelo escolhido e justificativa

O modelo escolhido foi o Random Forest, por virtude do fato de que apresenta resultados superiores aos demais.

4.5. Avaliação

Descreva a solução final de modelo preditivo e justifique a escolha. Alinhe sua justificativa com a Seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Descreva também um plano de contingência para os casos em que o modelo falhar em suas previsões.

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

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

ignorar

A solução final do modelo preditivo desenvolvido foi utilizar dois modelos separados, um treinado utilizando apenas pessoas que foram tratadas com terapia adjuvante e outro treinado apenas com pessoas que fizeram terapia neoadjuvante. Essa separação foi feita para que fosse possível gerar 2 modelos extremamente especializados em cada tratamento, visando ensiná-los quais casos funcionaram e quais não funcionaram em cada tipo de terapia (eles seguem a mesma lógica).

Aplicamos uma lógica que define o que era um caso de sucesso e o que era um caso de insucesso. Para determinar em qual cenário cada paciente se encaixava, aplicamos uma fórmula matemática baseada nas individualidades/dados clínicos de cada paciente. Para exemplificar, primeiro iremos analisar a gravidade da situação do paciente. Definimos uma variável chamada *info_iniciais_subtipo* e *info_iniciais_intensidade*, ambas possuem um valor pré definido de 0.15. (esse valor é dinâmico vai se alterando conforme o código vai analisando o caso do paciente).

```
def logica_target(df,i):
    #info iniciais
    info_iniciais_subtipo = 0.15
    if (df.loc[i, 'Tamanho do tumor'] == 1):
        info_iniciais_subtipo *= 0.1
    elif (df.loc[i, 'Tamanho do tumor'] == 2):
        info_iniciais_subtipo *= 0.3
    elif (df.loc[i, 'Tamanho do tumor'] == 3):
        info_iniciais_subtipo *= 0.6
    elif (df.loc[i, 'Tamanho do tumor'] == 4):
        info_iniciais_subtipo *= 1
    elif(df.loc[i, 'Tamanho do tumor'] == 0):
        info_iniciais_subtipo *= 0.01

    #calculando o fator de intensidade para a gravidade do tumor da paciente
    info_iniciais_intensidade = 0.15
    if (df.loc[i, 'Intensidade do tumor'] == 1):
        info_iniciais_intensidade *= 0.1
    elif (df.loc[i, 'Intensidade do tumor'] == 2):
        info_iniciais_intensidade *= 0.3
    elif (df.loc[i, 'Intensidade do tumor'] == 3):
        info_iniciais_intensidade *= 0.6
    elif (df.loc[i, 'Intensidade do tumor'] == 4):
        info_iniciais_intensidade *= 1
    elif(df.loc[i, 'Intensidade do tumor'] == 0):
        info_iniciais_subtipo *= 0.01
```

Se um paciente tiver Estadio Clínico 3A, por exemplo , a variável *info_iniciais_subtipo* será multiplicada por 0,6 porque o número 3 representa o tamanho do tumor, enquanto a variável *info_iniciais_intensidade* será multiplicada por 0,1, já que a letra A representa 1.

Após esse cálculo, iremos definir a gravidade por meio da soma dessas duas variáveis:

```
gravidade = info_iniciais_subtipo + info_iniciais_intensidade
```

Com essa informação em mãos, para o treinamento do modelo, nós analisamos a última informação do paciente (se ele sobreviveu ao tratamento sem sequelas, se sobreviveu mas continuou com câncer/teve recidiva e se ele faleceu por fatores cancerígenos). Por ser uma informação crucial para saber se o tratamento deu certo, decidimos novamente transformar essa informação em números para que seja possível analisar o quão bem sucedido foi o tratamento, concluímos isso depois de refletirmos que uma pessoa curada do câncer de mama sem complicações não é representa a mesma coisa que uma pessoa que também se curou mas teve recidivas ainda mais sérias. Baseando-se nessa ideia, criamos a variável *ultima_info_paciente*.

Nos casos em que o paciente sobreviveu sem outras especificações (VIVO S.O.E), nós definimos como sucesso

```
#calculando o fator final para mensurar o sucesso ou insucesso dependendo do estado atual e/ou recidiva e/ou tempo de sobrevida
ultima_info_paciente = 1
if (df.loc[i,'Última informação do paciente'] == "Vivo, SOE"):
    return 1
```

Para as pessoas que fizeram o tratamento mas continuaram com câncer, determinamos valores a serem assumidos para cada tipo de recidiva que o paciente possa ter desenvolvido, esse números foram definidos com base na severidade de cada tipo de recidiva (quanto mais distante, mais grave).

```
elif (df.loc[i,'Última informação do paciente'] == "Vivo, com câncer"):
    vivo_com_cancer = -0.6
    if (df.loc[i,'Recidiva Local'] == 1):
        recidiva_local = 0.4
        if(df.loc[i,'Tempo_ate_recidiva'] == '0 a 1 ano'):
            recidiva_local *= 1
        elif(df.loc[i,'Tempo_ate_recidiva'] == '1 a 3 anos'):
            recidiva_local *= 0.6
        elif(df.loc[i,'Tempo_ate_recidiva'] == '3 a 5 anos'):
            recidiva_local *= 0.2
        elif(df.loc[i,'Tempo_ate_recidiva'] == '5+ anos'):
            recidiva_local *= 0.05
        vivo_com_cancer *= recidiva_local

    elif(df.loc[i,'Recidiva Regional'] == 1):
        recidiva_regional = 0.7
        if(df.loc[i,'Tempo_ate_recidiva'] == '0 a 1 ano'):
            recidiva_regional *= 1
        elif(df.loc[i,'Tempo_ate_recidiva'] == '1 a 3 anos'):
            recidiva_regional *= 0.6
        elif(df.loc[i,'Tempo_ate_recidiva'] == '3 a 5 anos'):
            recidiva_regional *= 0.2
        elif(df.loc[i,'Tempo_ate_recidiva'] == '5+ anos'):
            recidiva_regional *= 0.05
        vivo_com_cancer *= recidiva_regional
```

Com a leitura do código, podemos perceber que recidiva local tem um peso menor do que recidiva regional, além disso também levamos em consideração o tempo até o surgimento da recidiva. Em contextos que a recidiva se manifestou pouco tempo após o tratamento, consideramos um caso mais crítico porque aparentemente a terapia feita não foi tão adequada, diferentemente de um contexto em que a pessoa só teve recidiva cancerígena 10 anos após o fim do tratamento. Após verificação do local da recidiva e do tempo até essa recidiva, os valores são multiplicados e o resultado é armazenado.

```
elif(df.loc[i,'Recidiva à distância'] == 1):
    recidiva_a_distancia = 0.9
    if(df.loc[i,'Tempo_ate_recidiva'] == '0 a 1 ano'):
        recidiva_a_distancia *= 1
    elif(df.loc[i,'Tempo_ate_recidiva'] == '1 a 3 anos'):
        recidiva_a_distancia *= 0.6
    elif(df.loc[i,'Tempo_ate_recidiva'] == '3 a 5 anos'):
        recidiva_a_distancia *= 0.2
```

Após a análise das pessoas que ficaram vivas com câncer, nós verificamos os casos de pessoas que morreram por conta do câncer, e por se tratar do pior cenário entre os outros 2 (Vivo, Vivo com câncer), definimos um peso mais rigoroso. Nessa situação, decidimos analisar quanto tempo de sobrevida o paciente teve, porque nos casos mais graves, quando a pessoa já está com seu destino definido, quanto mais tempo essa pessoa sobreviver, mais eficiente foi o tratamento.

Para simplificar a ideia, se o paciente morreu, mas teve grande tempo de sobrevida, o modelo tende a classificar como sucesso, mas se ele morreu com pouquíssimo tempo de sobrevida o modelo bem provavelmente vai classificá-lo como falha.

```
elif (df.loc[i, 'Última informação do paciente'] == 'Óbito por câncer'):
    morto_por_cancer = -0.75
    if(df.loc[i, 'Tempo_de_sobrevida'] == '0 a 1 ano'):
        morto_por_cancer *= 1
    elif(df.loc[i, 'Tempo_de_sobrevida'] == '1 a 3 anos'):
        morto_por_cancer *= 0.8
    elif(df.loc[i, 'Tempo_de_sobrevida'] == '3 a 5 anos'):
        morto_por_cancer *= 0.5
    elif(df.loc[i, 'Tempo_de_sobrevida'] == '5+ anos'):
        morto_por_cancer *= 0.1
```

Depois de todo esse cálculo finalmente chegamos ao score definido pelo código.

```
score = gravidade + ultima_info_paciente

if score > 0:
    return 1
else:
    return 0
```

Com esse score definimos nosso target e depois colocamos a base de dados num modelo de Random Forest para ser classificado como sucesso ou insucesso.

Acreditamos que esse raciocínio é válido porque ele analisa a gravidade dos casos, e com base no score gerado pela fórmula, é possível ter, de maneira numérica, uma predição do tratamento adequado para cada caso específico de tumor. Além disso, foram levados em consideração as indicações feitas pelo cliente do que determina ser um caso de sucesso e um caso de insucesso, dessa forma, tivemos informações de quais cenários são piores que outros (uma recidiva à distância é pior que uma recidiva local, por exemplo) fazendo com que essa diferença fosse analisada, impactando de maneira direta na resposta do modelo.

Nosso modelo preditivo é capaz de mostrar quais informações são mais relevantes para definir qual tratamento é o melhor, fazendo com que seja possível analisar o que interfere na taxa de sucesso de cada terapia. Outro ponto forte que atende ao entendimento de negócio é o fato de dados vazios terem sido tratados e os casos fora da curva (outliers) terem sido excluídos para que não interfiram na acurácia do modelo, ademais, evitamos ao máximo envolver vieses humanos inconscientes nos dados para que isso não prejudique seu desempenho

Por se tratar de um modelo baseado em evidências fundamentadas em dados clínicos, nosso modelo soluciona uma das maiores dores do nosso cliente: Não saber para qual tratamento encaminhar o paciente. Além disso, métodos inovadores como um modelo preditivo fazem com que a ameaça de novos produtos seja irrelevante, já que se trata de um artifício tecnológico.

Em caso de falha das previsões, o médico deve entrar em contato com o Inteli, para reportar imediatamente os erros observados durante a execução do modelo preditivo. Essa informação será repassada à orientadora responsável pelo projeto, que requisitará ao grupo de desenvolvedores do grupo uma correção dos problemas o mais rápido possível.

explicar lógica

- Por se tratar de um modelo baseado em evidências fundamentadas em dados clínicos, nosso modelo soluciona uma das maiores dores do nosso cliente que é a de não saber para qual tratamento encaminhar o paciente. Além disso, métodos inovadores como um modelo preditivo fazem com que a ameaça de novos produtos seja irrelevante, já que se trata de um artifício tecnológico. [] Nosso modelo é capaz de mostrar quais informações são mais relevantes para definir qual tratamento é o melhor, fazendo com que seja possível analisar o que interfere na taxa de sucesso de cada terapia. Outro ponto forte que atende ao entendimento de negócio é o fato de dados vazios terem sido tratados e os casos fora da curva (outliers) terem sido excluídos para que não interfiram na acurácia do modelo, ademais, evitamos ao máximo envolver vieses humanos inconscientes nos dados para que isso não prejudique seu desempenho. (relação com negocio
- *o pq isso faz sentido e suas peculiaridades*
- *relacionar com entendimento do negocio*
- *como isso facilita a vida do medico*
- *plano de contingenci*
- *explicabilidade*
- *verificar aceitação de hipoteses*
- **usar graficos e tabelas para enriquecer 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.

-resultados do projeto

-como o parceiro deve utilizar o modelo

- recomendações para ajudar parceiro de como tratar as pessoas afetadas (apens pacientes ou familiares e etc.?)

O modelo escolhido pelo grupo foi Random Forest, por ter uma acurácia maior - 92% para a terapia Adjuvante e de 91% para a terapia Neoadjuvante.

Apesar disso, ainda podem ocorrer recomendações incorretas, por isso é extremamente recomendado que seja realizada uma validação médica aprofundada para verificar os resultados, de forma que o produto seja autorizado a ser aplicado como ferramenta médica.

O uso da plataforma ocorre de maneira muito intuitiva, e por isso não há um manual de instruções. Nela, o usuário poderá obter a predição de dois modos: direto, onde ele preencherá um formulário com as informações daquela paciente e obtém em seguida o resultado; ou indireto, onde ele faz o upload de um arquivo .csv ou .xlsx (excel) para que o modelo faça a predição de cada uma das linhas na tabela de uma vez, retornando o resultado de uma vez e ainda gera uma tabela, onde é possível ver o resultado dos dois tratamentos em cada paciente.

A plataforma foi disponibilizada em um [site](#), mas caso seja desejado, é possível que o produto seja utilizado sem conexão com a internet. Para isso, basta baixar os arquivos no [repositório do projeto](#), fazer as instalações de algumas bibliotecas de python utilizadas no código e executar o código. Para instalar as bibliotecas, basta acessar o terminal da máquina local e inserir `pip install <nome_da_biblioteca>`, execute o comando acima para cada uma das bibliotecas a seguir: “joblib”, “PIL” e “streamlit”. Além disso, execute o seguinte comando: “pip install -U scikit-learn”. Assim, a máquina será capaz de rodar o produto livremente.

Como recomendações de uso, vale frisar o conselho dado pelo Dr. Roger Chamas de que, dada a clareza das informações de cada tratamento, é altamente recomendado que as

informações sejam apresentadas à paciente, de forma que ela possa escolher o próprio tratamento de forma ainda mais consciente, garantindo a transparência, e facilitando a escolha.

6. Referências

G1 SP (São Paulo). **Com mais de mil pacientes com câncer à espera de cirurgia, governo de SP anuncia 45 leitos e 3 salas cirúrgicas na tentativa de reduzir fila:** segundo a secretaria estadual da saúde, meta da gestão é zerar fila nos 100 primeiros dias do ano; haverá também a ativação de 393 leitos ociosos no hospital das clínicas da faculdade de medicina da usp.. Segundo a Secretaria Estadual da Saúde, meta da gestão é zerar fila nos 100 primeiros dias do ano; haverá também a ativação de 393 leitos ociosos no Hospital das Clínicas da Faculdade de Medicina da USP.. 2023. Disponível em: <https://g1.globo.com/sp/sao-paulo/noticia/2023/01/24/com-mais-de-mil-pacientes-com-cancer-a-espera-de-cirurgia-governo-de-sp-anuncia-45-leitos-e-3-salas-cirurgicas-na-tentativa-de-reduzir-fila.ghtml>. Acesso em: 23 fev. 2023.

GIGLIO, Auro del. **ONCOLOGISTA DO HCOR APONTA 10 DICAS PARA PREVENÇÃO DO CÂNCER:** a prevenção dos diversos tipos de câncer inclui, basicamente, a adoção de uma vida saudável, com alimentos que previnem o câncer e atividades físicas.. A prevenção dos diversos tipos de câncer inclui, basicamente, a adoção de uma vida saudável, com alimentos que previnem o câncer e atividades físicas.. 2021. Disponível em: https://www.hcor.com.br/imprensa/noticias/oncologista-do-hcor-aponta-10-dicas-para-prevencao-do-cancer/?gclid=CjwKCAiAioifBhAXEiwApzCztpKeXJbn6tunOQIO8T6Cawb40AZJ6SFccPqH2riiD_Gx1Moi2MEvoBoCQoUQAvD_BwE. Acesso em: 23 fev. 2023.

GOVERNO, Do Portal do. **Instituto do Câncer de São Paulo recebe selo de reacreditação internacional:** icesp foi o primeiro hospital da rede pública da capital a ser acreditado pela joint commission international (jci), em 2014. Icesp foi o primeiro hospital da rede pública da capital a ser acreditado pela Joint Commission International (JCI), em 2014. 2021. Disponível em:

<https://www.saopaulo.sp.gov.br/spnoticias/orgaos-governamentais/secretaria-da-saude/instituto-do-cancer-de-sao-paulo-recebe-selo-de-reacreditacao-internacional/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **Instituto do Câncer do Estado de São Paulo.** 2022. Disponível em: <https://icesp.org.br/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **Mitos e Verdades Sobre o Câncer.** 2022. Disponível em: <https://icesp.org.br/mitos-e-verdades-sobre-o-cancer/>. Acesso em: 23 fev. 2023.

ICESP (São Paulo). **RESIDENTES DA ONCOLOGIA CLÍNICA DO ICESP OBTÊM MÉDIA MAIS ALTA EM EXAME MUNDIAL.** 2023. Disponível em: <https://icesp.org.br/noticias/residentes-da-oncologia-clinica-do-icesp-obtem-media-mais-alta-em-exame-mundial/>. Acesso em: 23 fev. 2023.

INCA (Rio de Janeiro). **Detecção Precoce do Câncer.** 2021. Disponível em: <https://www.inca.gov.br/sites/ufu.sti.inca.local/files/media/document/deteccao-precoce-do-cancer.pdf>. Acesso em: 23 fev. 2023.

MORSCH, José Aldair. **TEMPO DE GUARDA DE PRONTUÁRIO MÉDICO: VEJA QUAL É O PRAZO E COMO SE ORGANIZAR.** 2022. Disponível em: <https://telemedicinamorsch.com.br/blog/tempo-de-guarda-de-prontuario-medico#:~:text=O%20tempo%20de%20guarda%20de%20prontu%C3%A1rio%20m%C3%A9dico%20no%20Brasil%20corresponde,Em%20seu%20Art.> Acesso em: 23 fev. 2023.

O ESTADO DE S.PAULO (São Paulo). **Acesso a novos tratamentos pelo SUS ainda é um obstáculo:** drogas mais modernas têm alto custo, e a maioria não está disponível no sistema público. Drogas mais modernas têm alto custo, e a maioria não está disponível no sistema público. 2019. Disponível em:

<https://www.anahp.com.br/noticias/acesso-a-novos-tratamentos-pelo-sus-ainda-e-um-obstaculo/>. Acesso em: 23 fev. 2023.

Anexos

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