



NeoVision USP Medicina

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 | 4 |
| 2. Objetivos e Justificativa | 5 |
| 2.1. Objetivos | 6 |
| 2.2. Proposta de Solução | 6 |
| 2.3. Justificativa | 6 |
| 3. Metodologia | 6 |
| 4. Desenvolvimento e Resultados | 7 |
| 4.1. Compreensão do Problema | 7 |
| 4.1.1. Contexto da indústria | 7 |
| 4.1.2. Análise SWOT | 10 |
| 4.1.3. Planejamento Geral da Solução | 12 |
| 4.1.4. Value Proposition Canvas | 13 |
| 4.1.5. Matriz de Riscos | 13 |
| 4.1.6. Personas | 14 |
| 4.1.7. Jornadas do Usuário | 16 |
| 4.1.8. <i>User Stories</i> | 17 |
| 4.1.9. Política de Privacidade | 19 |
| 4.2. Compreensão dos Dados | 20 |
| 4.3. Preparação dos Dados e Modelagem | 21 |
| 4.4. Comparação de Modelos | 22 |
| 4.5. Avaliação | 23 |
| 5. Conclusões e Recomendações | 24 |
| 6. Referências | 25 |
| Anexos | 26 |

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

Cross Industry Standard Process for Data Mining, o CRISP-DM é um método ágil para planejamento e desenvolvimento de *machine learning*. Ele é separado em 6 etapas, desde o entendimento do negócio até a implantação:

- **Entendimento do negócio:** nessa primeira etapa, a meta é entender o problema de negócio que se deseja resolver com a mineração de dados e definir os objetivos do projeto. Este passo é importante para entender as dores do cliente, o que eles te oferecem e como você irá trabalhar daqui pra frente.
- **Entendimento dos dados:** envolve a coleta, exploração e mineração dos dados. Esse estágio é extremamente importante para que haja a familiarização com os dados, garantindo fidedignidade, qualidade e relevância. O objetivo é coletar e explorar os dados disponíveis, avaliar sua qualidade e ver se eles são adequados para o problema em questão.
- **Preparando os dados:** aqui, o objetivo é preparar os dados para análise, incluindo limpeza, transformação, integração e seleção de variáveis. Dados de entradas ruins resultam em dados de saídas ruins, logo é importante dar como entrada os dados corretos.
- **Modelagem:** neste estágio, envolve técnicas e algoritmos de machine learning, avaliando qual melhor se adequa aos objetivos do seu projeto. É recomendado separar o conjunto de dados entre treino e teste. No de treino serão gerados os modelos, e no de teste será a parte de validação do modelo.
- **Avaliação:** Um passo antes da implementação, aqui o objetivo é avaliar a qualidade e eficácia dos modelos criados, verificar se eles atendem aos objetivos do projeto e determinar se é necessário ajustes, levando em consideração os objetivos iniciais do primeiro estágio, de tal forma que os modelos propostos consigam atender os objetivos pré definidos.
- **Implantação:** sendo o último passo da metodologia, sua importância se encontra em colocar os modelos em produção e monitorar o seu desempenho ao longo do tempo, garantindo que ele atenda aos objetivos do projeto e continue a ser eficaz.

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.

4.1.1.2. 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:
 - o 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.
 - o 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

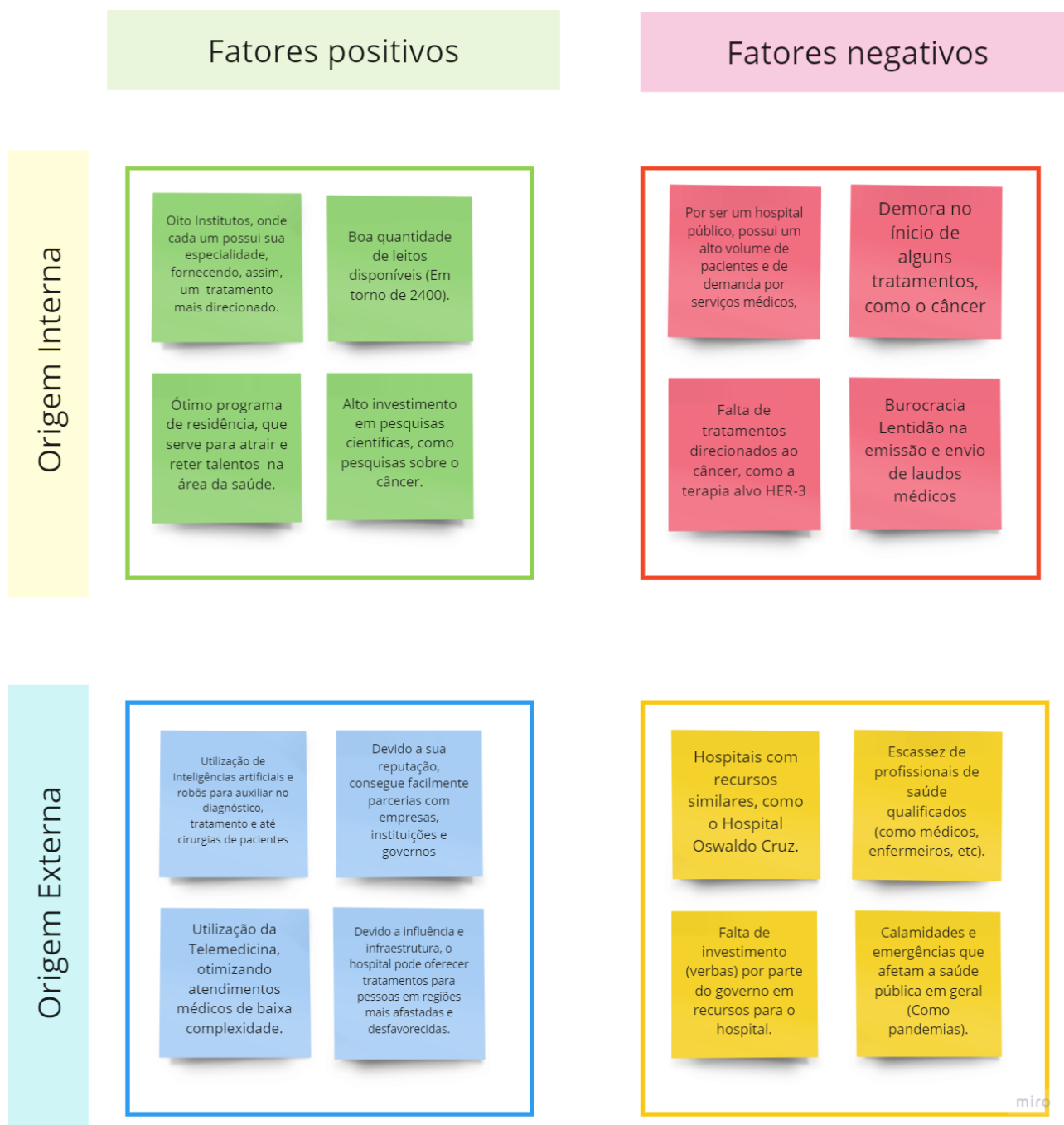
A análise SWOT é uma técnica de planejamento estratégico amplamente utilizada em empresas, organizações e projetos. A sigla SWOT significa Strengths (Forças), Weaknesses (Fraquezas), Opportunities (Oportunidades) e Threats (Ameaças). Nesse contexto, ela visa identificar as forças e fraquezas internas 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.

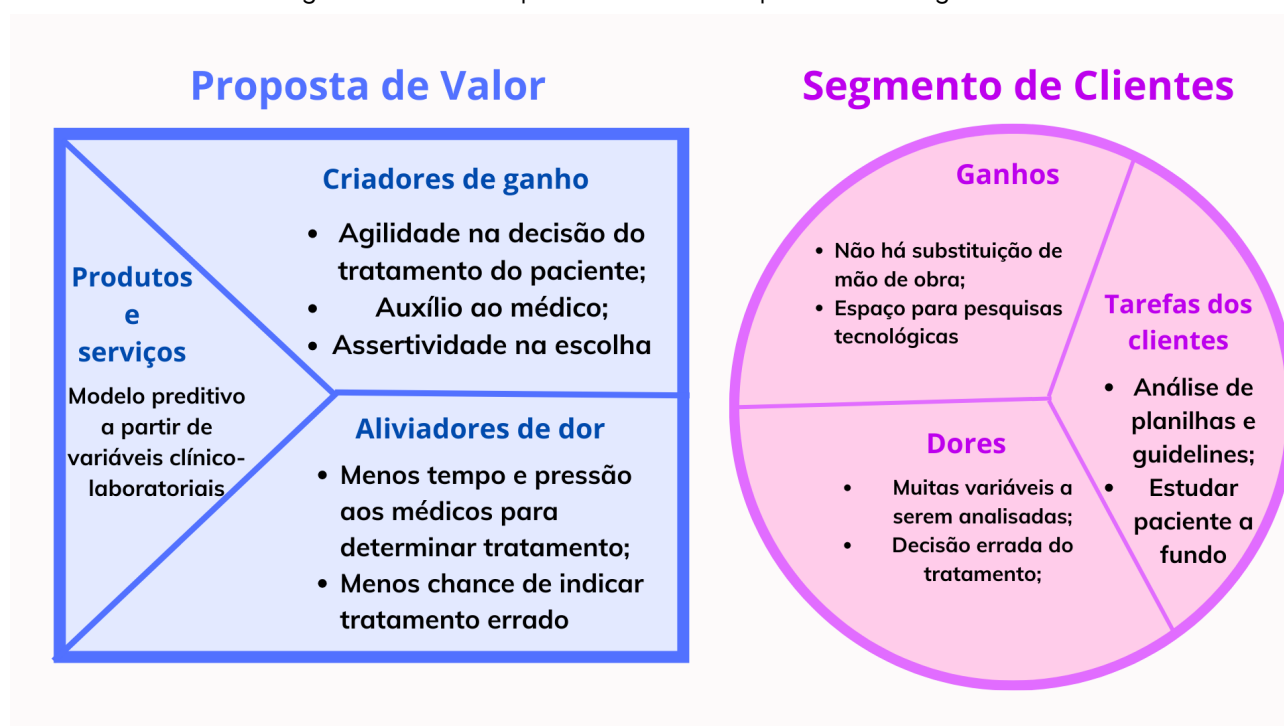
4.1.3.5. Qual será o critério de sucesso e qual medida será utilizada para o avaliar

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 interpretação 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 sobrevivência 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

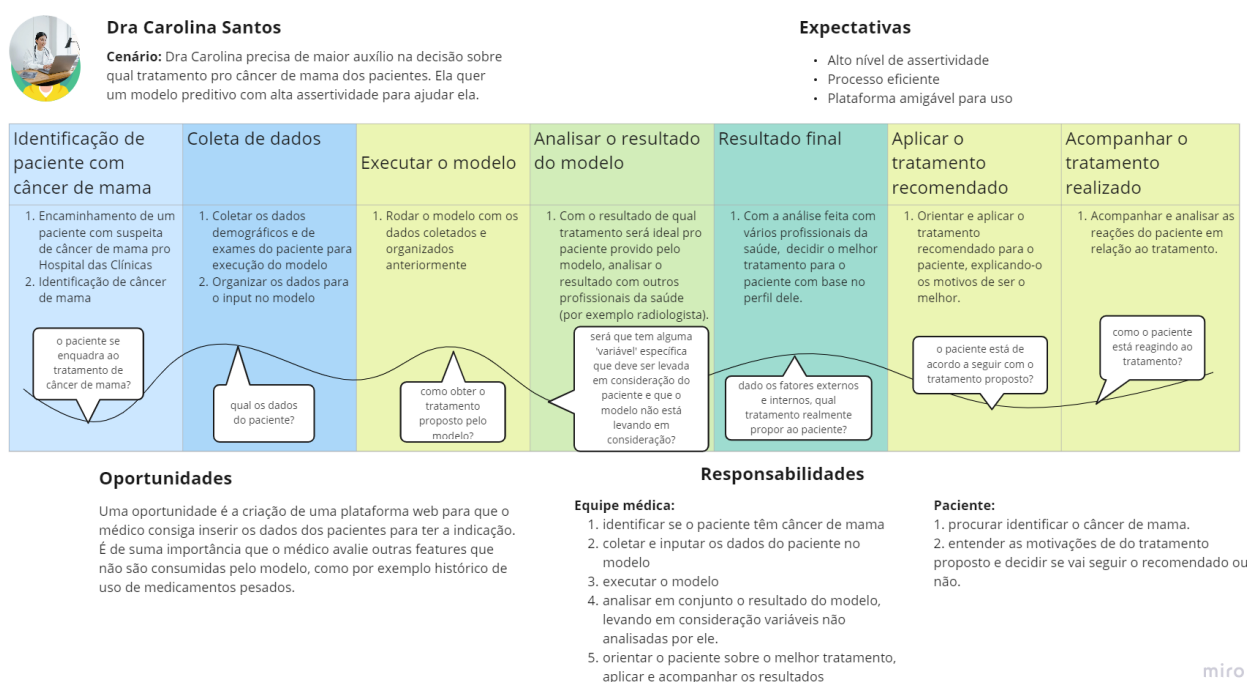
Um mapa de jornada é uma visualização do processo pelo qual uma pessoa passa para atingir um objetivo. O mapeamento da jornada começa compilando uma série de ações do usuário em uma linha do tempo. Em seguida, a linha do tempo é desenvolvida com pensamentos e emoções do usuário para criar uma narrativa.

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

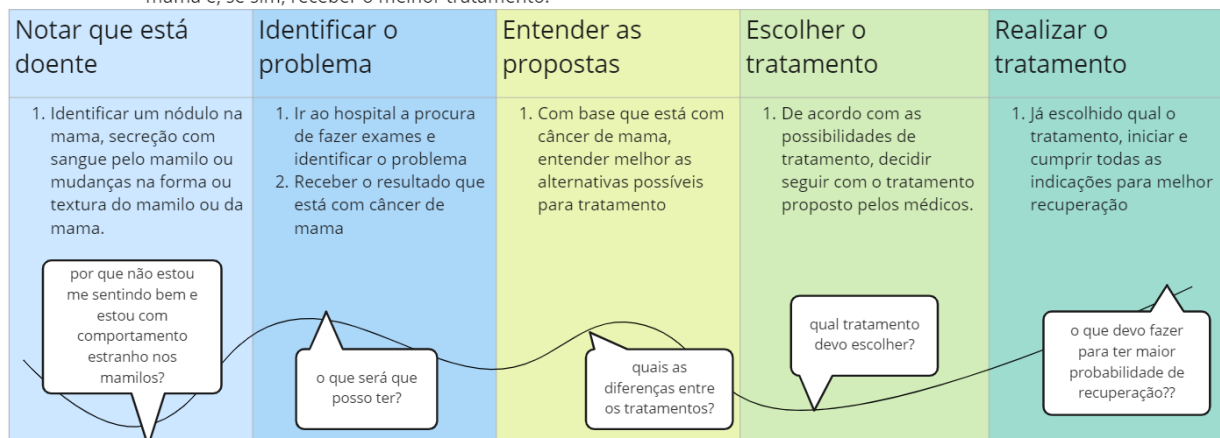


Fernanda Silveira

Cenário: Fernanda não está se sentindo bem, devido ao histórico familiar quer confirmar se está com câncer de mama e, se sim, receber o melhor tratamento.

Expectativas

Receber o melhor tratamento de acordo com o seu perfil, que lhe dê maiores chances de recuperação



Oportunidades

1. É necessário acesso facilitado aos exames para identificação do câncer de mama
2. É preciso acesso facilitado e explicitado os fatores para a escolha do melhor tratamento da Fernanda.

Responsabilidades

Paciente:

1. procurar identificar o câncer de mama
2. entender as motivações de do tratamento proposto e decidir se vai seguir o recomendado ou não.

Equipe médica:

1. identificar se o paciente têm câncer de mama
2. coletar os dados do paciente
3. analisar em conjunto os dados do paciente e utilizar uma ferramenta de IA para maior assertividade na proposta do modelo, levando em consideração variáveis não analisadas por ele.
4. orientar o paciente sobre o melhor tratamento, aplicar e acompanhar os resultados

miro

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

1. Informações gerais sobre a empresa/organização;

O NeoVision é um grupo de estudantes do Instituto de Tecnologia e Liderança (INTELI) trabalhando em conjunto para desenvolver um modelo preditivo para ajudar os médicos a indicar o melhor tratamento para o câncer de mama. O projeto é parte do terceiro módulo da faculdade e conta com membros qualificados que combinam seus conhecimentos para criar um modelo confiável e preciso. O NeoVision está comprometido em usar a tecnologia para otimizar a indicação do tratamento mais eficaz e acelerar o processo de tomada de decisão para o tratamento de pacientes com câncer de mama.

2. Quais dados pessoais são coletados

Os dados coletados são em sua maioria dados clínicos como, por exemplo, histórico médico, considerados sensíveis. Além disso, nosso modelo preditivo usa dados demográficos e antropométricos (como peso e altura), bem como possivelmente informações para contato/registro como e-mail e telefone.

3. Onde os dados são coletados (fonte);

Os dados são coletados de bases de dados disponibilizadas por pesquisadores da Faculdade de Medicina da Universidade de São Paulo e do Instituto do Câncer do Estado de São Paulo.

4. Para quais finalidades os dados são utilizados

Os dados coletados serão utilizados apenas para a decisão entre os tratamentos adjuvante e neoadjuvante para pacientes com câncer de mama. Ademais, os dados podem permanecer armazenados no sistema para futuras avaliações em estudos clínicos.

5. Onde os dados ficam armazenados;

Os dados coletados estão armazenados em um servidor seguro, pois é muito importante assegurar que as informações estão protegidas contra acesso não autorizado, perda de dados e outros riscos.

6. Qual o período de armazenamento dos dados (retenção)

O período de armazenamento dos dados é indeterminado, podendo ser removido por solicitação do titular a qualquer momento.

7. Uso de cookies e/ou tecnologias semelhantes

Não faremos uso de cookies ou tecnologias de rastreamento semelhantes.

8. Com quem esses dados são compartilhados (parceiros, fornecedores, subcontratados)

Esses dados são compartilhados somente com a equipe da Neovision e seus parceiros: o Inteli e a Faculdade de Medicina da USP.

9. Informações sobre medidas de segurança adotadas pela empresa;

Todos os dados são prontamente anonimizados e o acesso a eles é restrito, armazenados criptograficamente em um servidor na nuvem.

10. - Orientações sobre como a empresa/organização atende aos direitos dos usuários

Nosso grupo atende aos direitos dos usuários com as seguintes medidas: implementando uma política de segurança de dados, sendo transparente em relação a todo o tratamento de dados e mostrando prontidão para responder às solicitações referentes aos direitos dos usuários.

11. Informações sobre como o titular de dados pode solicitar e exercer os seus direitos;

O titular pode exercer os seus direitos fazendo uma solicitação ao controlador de dados para receber uma cópia de todos os dados que possuem sobre ele, ou solicitar que certos dados sejam corrigidos ou excluídos. Ademais, o titular tem direito de receber

informações sobre como e por que seus dados são usados, e para onde eles são transferidos.

12. Informações de contato do Data Protection Officer (DPO) ou encarregado de proteção de dados da organização.

Caso o titular tenha alguma dúvida ou deseja exercer algum direito sobre seus dados, ele pode entrar em contato conosco através do e-mail 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 |
|--|----------------|
| Receptor de Estrogênio (quantificação %) | categórico |
| Índice H (Receptor de 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 grávida? | 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 menstruação | 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 centímetros) | 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. |

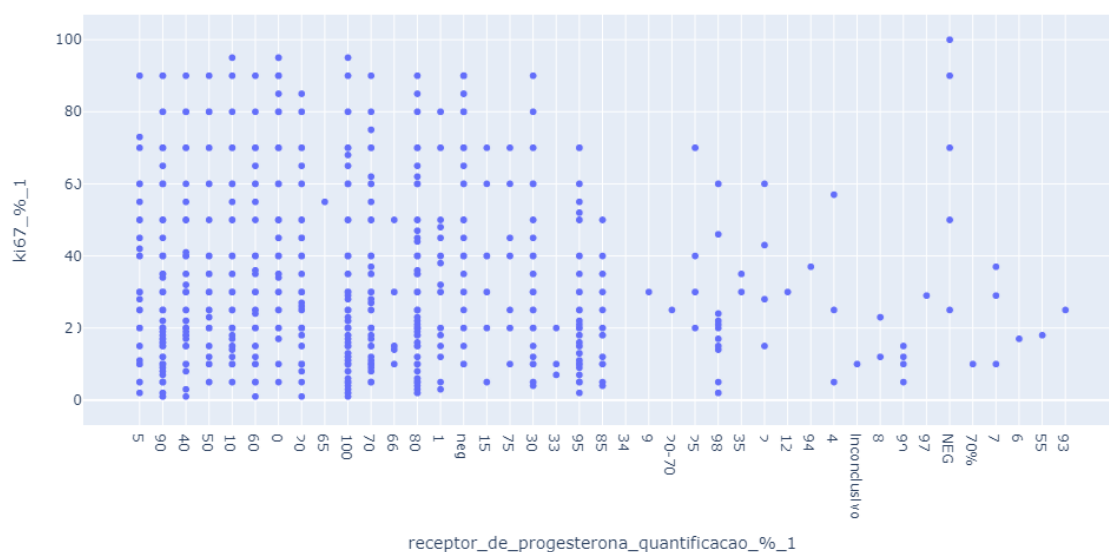
| TABELA PESO E ALTURA | |
|----------------------|---|
| Colunas numéricas | clique aqui para acessar a célula no colab. |

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 %)'

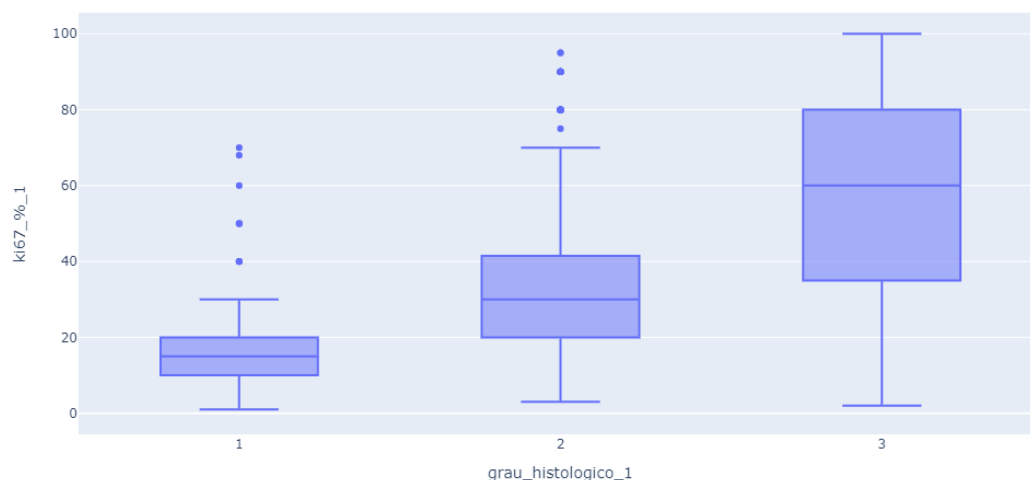


Fonte: Elaboração dos autores.

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(%).

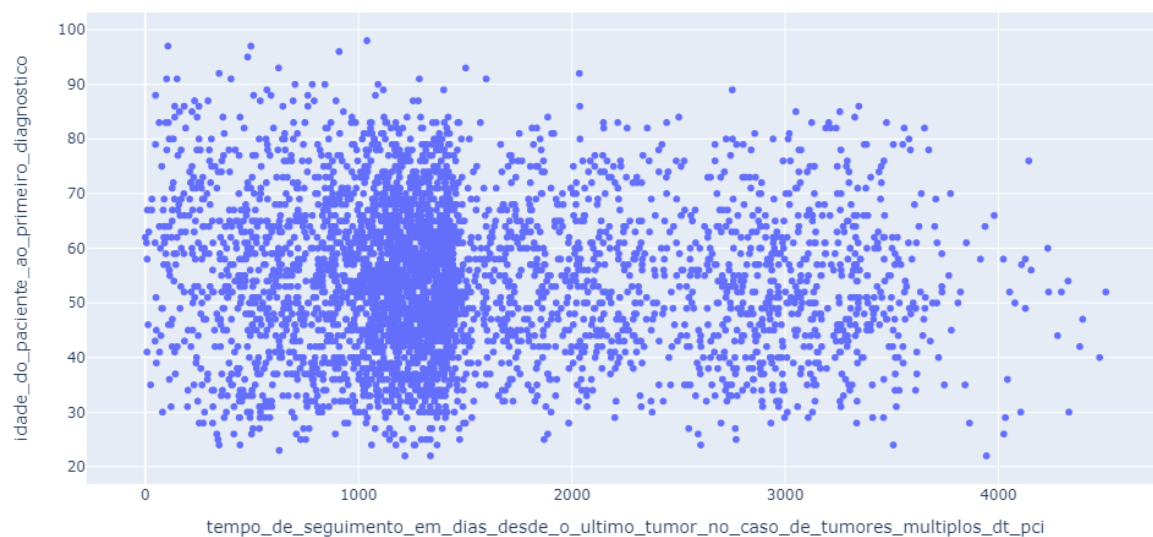
Relação entre variáveis na tabela de demográfico

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'

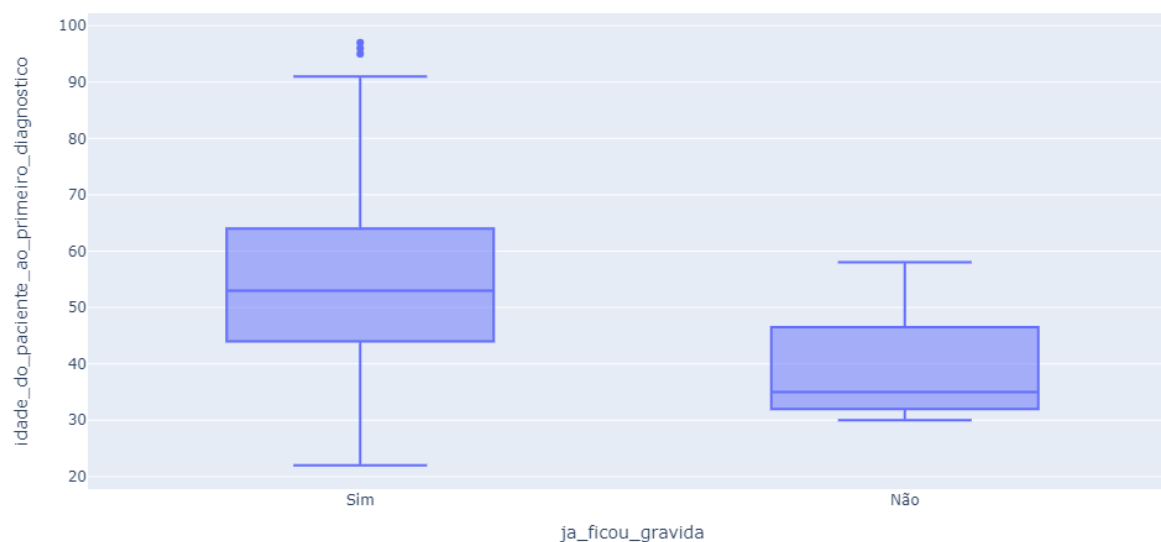


Fonte: Elaboração dos autores.

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'



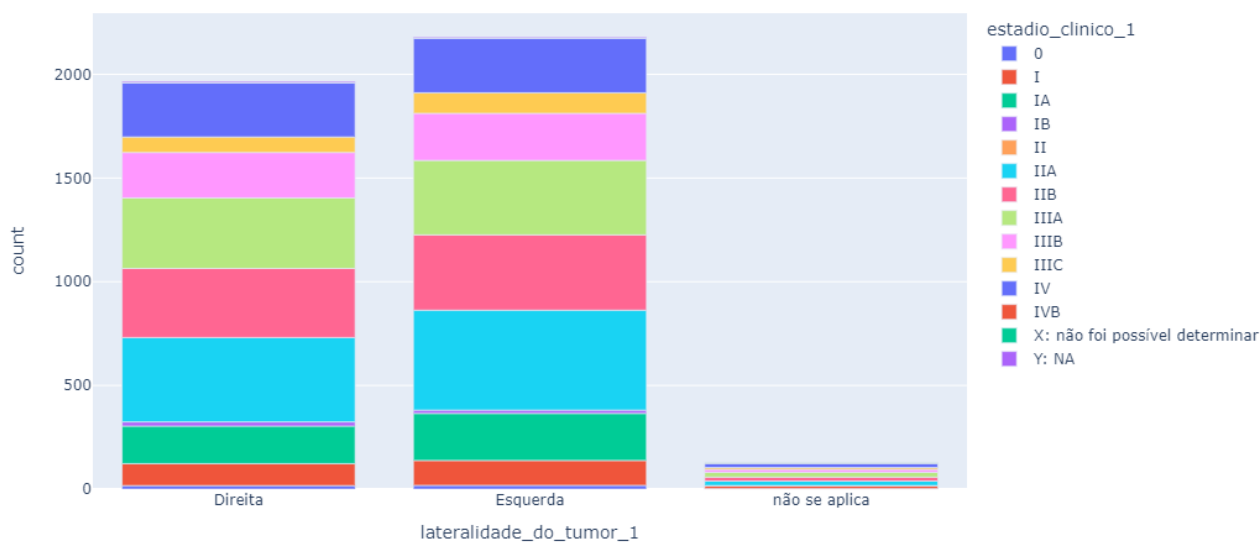
Fonte: Elaboração dos autores.

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'

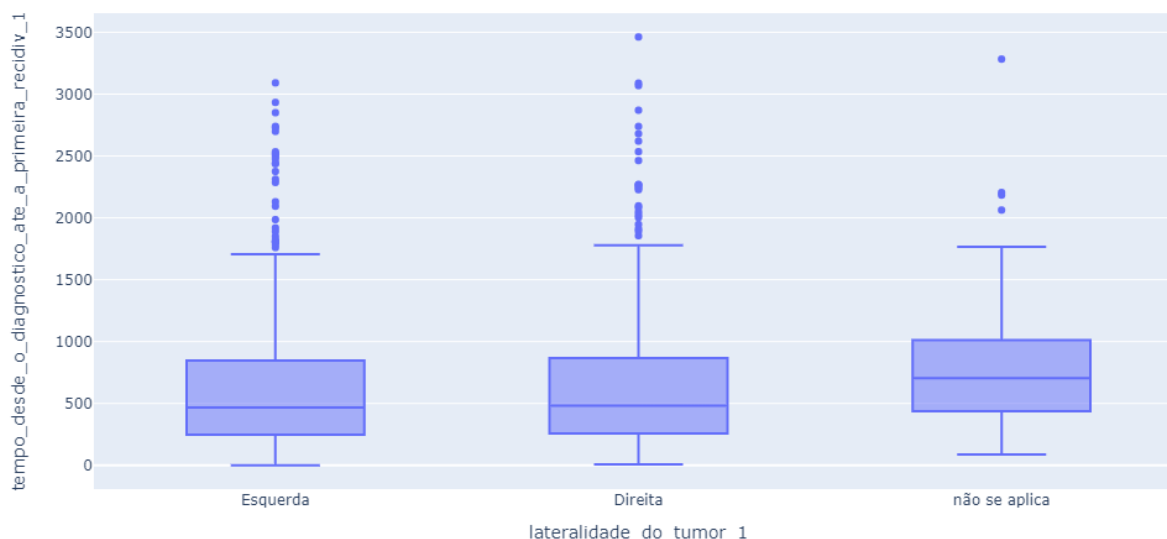


Fonte: Elaboração dos autores.

Através do gráfico anterior é possível perceber que a proporção 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_instances()' 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, 2977.0, ...] |
| 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, 97.0, ...] |
| 13 | tempo_de_seguimento_(em_dias)__desde_o_último... | [] | [4153.0, 4330.0, 4381.0, 3734.0, 4474.0, 3864.0, ...] |
| 14 | quantas_vezes_ficou_grávida | [] | [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, 53.0, ...] |
| 17 | por_quanto_tempo_amamentou | [] | [82.0, 100.0, 178.0, 84.0, 96.0, 240.0, 150.0, ...] |
| 18 | idade_da_primeira_menstruacao | [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_grávida', 'idade_na_primeira_gestacao', 'por_quanto_tempo_amamentou', 'idade_da_primeira_menstruacao']
```

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).

Proposta de features

Visto que nossa problemática é voltada à predição de dois tipos de tratamento, nosso projeto se encaixa como Modelo Supervisionado, onde conhecemos as entradas e saídas dos dados.

As features apresentadas neste modelo preditivo para tratamento de câncer de mama possuem diferentes intenções e impactos na predição do tipo de tratamento necessário. É notável a importância que as características tanto hormonais e tumorais quanto hábitos de vida influenciam no projeto.

As variáveis hormonais, como **recep_progesterona_qtd** e **recep_estrogenio_qtd**, têm o objetivo de identificar a presença de receptores hormonais no tumor, o que pode ser um fator importante para definir o tipo de tratamento. A atividade do gene HER2, medida pela variável **her2_qtd**, é um indicador importante para a efetividade da terapia alvo, trazendo um importante impacto na escolha também. A idade e a presença de atividade física, como mencionado, podem ser importantes indicadores de risco, assim como o histórico familiar, representado pela variável **historia_familiar_de_cancer_relacionado**. Tendo isso em vista, aqui estão as nossas features e suas explicações:

- **tipo_histológico**: se refere ao tipo de tecido que o tumor é formado, por exemplo, carcinoma de células escamosas ou adenocarcinoma.
- **grau_histológico**: classifica a aparência das células cancerígenas em relação às células normais do tecido, geralmente sendo classificadas em graus 1, 2 e 3.
- **subtipo_tumoral**: pode ser uma subdivisão do tipo histológico, indicando características específicas do tumor.
- **recep_progesterona_qtd** e **recep_estrogênio_qtd**: se referem à quantidade de receptores de progesterona e estrogênio presentes nas células tumorais. Esses receptores podem influenciar no tratamento e prognóstico do câncer.
- **ki67_qtd**: é uma medida da taxa de proliferação celular e pode ser usada para prever o quão rápido um tumor está crescendo.
- **HER2_qtd**: é uma proteína presente em algumas células cancerígenas e sua presença pode indicar um tipo de câncer mais agressivo.
- **grupo_idade**: é a idade da pessoa no momento do diagnóstico.
- **já_ficou_grávida** e **idade_na_primeira_gestação**: indicam informações sobre a história reprodutiva da pessoa.

- **tempo_amamentação:** indica por quanto tempo a pessoa amamentou, o que pode afetar o risco de câncer de mama.
- **idade_da_primeira_menstruação:** é a idade em que a pessoa teve sua primeira menstruação
- **atividade_física, consumo_de_tabaco e consumo_de_álcool:** indicam hábitos de vida que podem estar relacionados ao risco de desenvolver câncer.
- **menopausa:** se a pessoa já entrou na menopausa ou não.
- **grupo_idade_na_primeira_gestação:** é a idade da pessoa quando teve sua primeira gestação.
- **história_familiar_de_câncer_relacionado e grau_parentesco_familiar_câncer:** indicam se há um histórico de câncer na família e qual o grau de parentesco dos familiares afetados.
- **código_da_topografia_cid_o e código_da_morfologia_de_acordo_com_o_cid_o:** são códigos que indicam a localização e a aparência do tumor, seguindo a classificação do CID-O (Classificação Internacional de Doenças para Oncologia).
- **tipo_histologico:** é o tipo de células que compõem o tumor, por exemplo, adenocarcinoma, carcinoma de células escamosas, etc.
- **classificacao_tnm_clinico_t_1:** é a classificação clínica do tamanho do tumor primário no momento do diagnóstico.
- **classificacao_tnm_clinico_n:** é a classificação clínica da presença e extensão de metástases nos linfonodos regionais.
- **classificacao_tnm_clinico_m:** é a classificação clínica da presença ou ausência de metástases distantes (por exemplo, nos pulmões ou no fígado).
- **lateralidade_do_tumor:** é a informação sobre em qual lado do corpo o tumor está localizado, por exemplo, esquerdo, direito ou bilateral.
- **classificacao_tnm_patologico_n:** é a classificação patológica da presença e extensão de metástases nos linfonodos regionais, determinada após a cirurgia.
- **classificacao_tnm_patologico_t:** é a classificação patológica do tamanho do tumor primário, determinada após a cirurgia.
- **possui_metastase:** informação sobre a presença ou ausência de metástases em algum outro local do corpo além do tumor primário.
- **risk_metastase:** informação sobre o risco de o tumor se espalhar para outros locais do corpo.
- **estagio_tumor:** é uma classificação geral do estágio do câncer, com base na extensão do tumor e na presença de metástases, que ajuda a orientar o tratamento.
- **imc:** índice de massa corporal do paciente, que é calculado dividindo o peso pela altura ao quadrado e é usado como uma medida de obesidade e estado nutricional.

b) Métricas relacionadas ao modelo (conjunto de testes, pelo menos 3).

Métricas de avaliação

Para avaliar a eficácia do nosso modelo preditivo, estamos utilizando três principais métricas: *precision*, *F1 Score* e *recall*. As três métricas foram baseadas na Matriz de Confusão, uma matriz 2x2 que avalia quatro tipos de valores em seu modelo: **Verdadeiro positivo**, quando o modelo prevê um valor positivo e ele realmente é; **Falso positivo**, quando ele prevê um valor positivo e ele não é; **Falso verdadeiro**, quando prevê falso e realmente é. **Falso negativo**, quando prevê um valor falso e ele não é. Sendo assim, surge os seguintes cálculos:

- **Acurácia:** A acurácia é uma métrica utilizada para medir a precisão geral de um modelo de classificação. Ela é calculada como a proporção de predições corretas em relação ao total de predições feitas pelo modelo.
- **Recall:** A métrica de recall avalia a proporção de positivos identificados corretamente. Para isso, seu cálculo é a razão entre verdadeiros positivos sobre a soma de verdadeiros positivos com negativos falsos.
- **F-1 Score:** É uma métrica que combina as medidas de precisão e recall para avaliar o desempenho de um modelo de classificação.

c) Apresentar o primeiro modelo candidato, e uma discussão sobre os resultados deste modelo (discussão sobre as métricas para esse modelo candidato).

Execução dos algoritmos de predição utilizando a biblioteca LazyPredict

Utilizando a biblioteca Lazy Predict, onde vários algoritmos são aplicados ao projeto, o modelo que mais se destacou foi o Random Forest. O Random Forest é um algoritmo de aprendizado de máquina que pode ser utilizado tanto para classificação quanto para regressão. Ele consiste em construir várias árvores de decisão e combiná-las para melhorar a precisão da predição.

As árvores de decisão são estruturas em forma de árvore que auxiliam na tomada de decisões baseadas em regras simples. Cada nó da árvore representa uma condição, como "se a idade é maior que 30 anos" ou "se a renda é menor que 50 mil". As decisões são tomadas seguindo os ramos da árvore até chegar a uma folha, que representa a classificação final.

Quando uma nova amostra é apresentada ao modelo Random Forest, as árvores individuais realizam suas previsões e a previsão final é baseada na votação da maioria das árvores. Por exemplo, se cinco árvores preveem que uma determinada amostra é da classe A e quatro preveem que é da classe B, o modelo Random Forest classifica a amostra como sendo da classe A.

| | | | | | |
|------------------------------------|------|------|------|------|------|
| AdaBoostClassifier | 0.79 | 0.77 | None | 0.79 | 0.47 |
| RandomForestClassifier | 0.78 | 0.76 | None | 0.78 | 0.38 |
| LGBMClassifier | 0.76 | 0.75 | None | 0.76 | 0.23 |
| ExtraTreesClassifier | 0.77 | 0.75 | None | 0.76 | 0.39 |
| NearestCentroid | 0.74 | 0.75 | None | 0.74 | 0.04 |
| BernoulliNB | 0.76 | 0.74 | None | 0.76 | 0.05 |
| SVC | 0.75 | 0.72 | None | 0.74 | 0.19 |
| NuSVC | 0.74 | 0.72 | None | 0.73 | 0.21 |
| LogisticRegression | 0.73 | 0.71 | None | 0.73 | 0.05 |
| BaggingClassifier | 0.74 | 0.71 | None | 0.73 | 0.26 |
| RidgeClassifierCV | 0.73 | 0.71 | None | 0.73 | 0.07 |
| LinearDiscriminantAnalysis | 0.73 | 0.71 | None | 0.73 | 0.07 |
| LinearSVC | 0.73 | 0.71 | None | 0.73 | 0.20 |
| SGDClassifier | 0.72 | 0.71 | None | 0.72 | 0.10 |
| RidgeClassifier | 0.73 | 0.71 | None | 0.73 | 0.05 |
| CalibratedClassifierCV | 0.73 | 0.70 | None | 0.72 | 1.03 |
| PassiveAggressiveClassifier | 0.70 | 0.67 | None | 0.69 | 0.05 |
| Perceptron | 0.68 | 0.66 | None | 0.68 | 0.05 |
| KNeighborsClassifier | 0.66 | 0.64 | None | 0.66 | 0.07 |
| DecisionTreeClassifier | 0.65 | 0.64 | None | 0.65 | 0.06 |
| ExtraTreeClassifier | 0.65 | 0.64 | None | 0.65 | 0.05 |
| LabelSpreading | 0.62 | 0.62 | None | 0.63 | 0.27 |

Fonte: Elaboração própria

De acordo com a tabela acima, pode-se observar que com o Random Forest, obtivemos uma acurácia de 76%.

Execução dos algoritmos de predição individualmente

Depois da execução com o Lazy Predict, o foco será na execução dos algoritmos de predição individualmente, com o objetivo de obter uma avaliação mais precisa do desempenho de cada

modelo. Ao utilizar a biblioteca LazyPredict para rodar vários algoritmos de predição de uma só vez, tivemos uma ideia geral do desempenho de cada modelo. No entanto, essa abordagem pode não fornecer informações detalhadas o suficiente sobre como cada modelo está se saindo em relação à tarefa específica de predição.

Ao rodar cada algoritmo de forma individual, é possível avaliar o desempenho de cada modelo em relação a métricas específicas, como precisão, recall, F1-score, acurácia, entre outras. Com isso, podemos comparar os resultados obtidos por cada modelo e identificar aquele que melhor se adapta à tarefa de predição em questão.

Além disso, essa abordagem permite uma análise mais detalhada dos pontos fortes e fracos de cada modelo, o que pode ajudar a identificar possíveis melhorias ou ajustes que podem ser feitos para otimizar o desempenho do modelo.

Os modelos que escolhemos para essa execução individual foram:

- K-Nearest Neighbors (KNN)
- Random Forest
- Naive Bayes
- Decision Tree
- Regressão Logística

A razão pela qual esses modelos foram escolhidos foi devido a sua popularidade em problemas relacionados a Machine Learning e também pelo seu desempenho, visto que esses algoritmos possuem uma grande abrangência e podem ser utilizados em diversos problemas de Machine Learning.

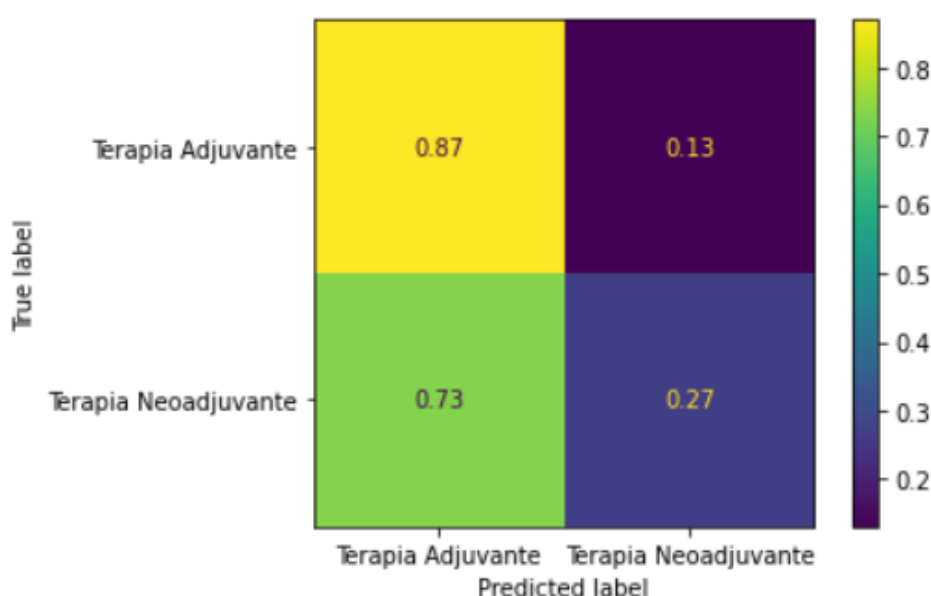
As métricas obtidas para cada um dos 5 modelos testados foram as seguintes:

- KNN:
 - Acurácia de treino: 66%
 - Acurácia de teste: 63%
 - Recall Terapia Adjuvante: 86%
 - Recall Terapia Neoadjuvante: 26%
 - F-1 Score Terapia Adjuvante: 74%
 - F-1 Score Terapia Neoadjuvante: 36%
- Random Forest:
 - Acurácia de treino: 100%
 - Acurácia de teste: 78%
 - Recall Terapia Adjuvante: 84%
 - Recall Terapia Neoadjuvante: 68%
 - F-1 Score Terapia Adjuvante: 82%
 - F-1 Score Terapia Neoadjuvante: 71%
- Naive Bayes:
 - Acurácia de treino: 74%

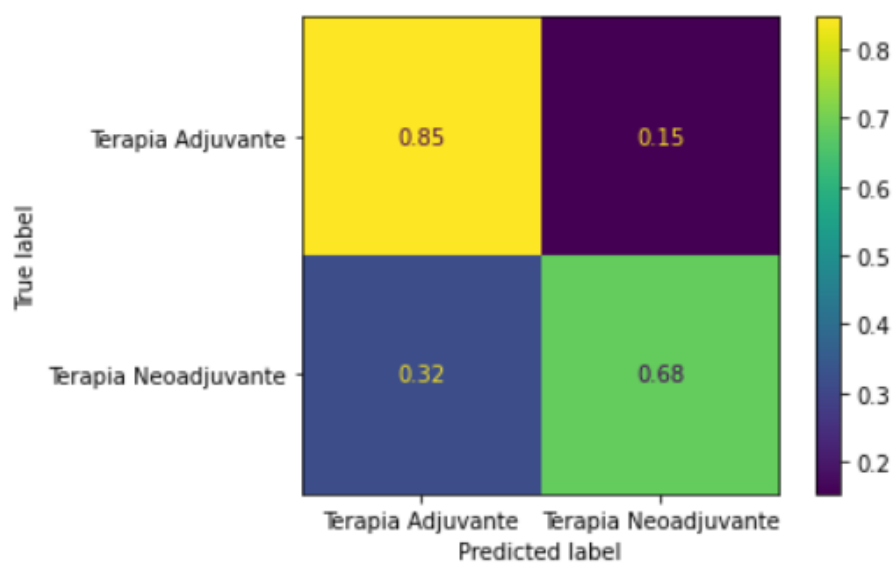
- Acurácia de teste: 71%
- Recall Terapia Adjuvante: 86%
- Recall Terapia Neoadjuvante: 46%
- F-1 Score Terapia Adjuvante: 78%
- F-1 Score Terapia Neoadjuvante: 55%
- Decision Tree:
 - Acurácia de treino: 75%
 - Acurácia de teste: 76%
 - Recall Terapia Adjuvante: 86%
 - Recall Terapia Neoadjuvante: 60%
 - F-1 Score Terapia Adjuvante: 81%
 - F-1 Score Terapia Neoadjuvante: 66%
- Regressão Logística:
 - Acurácia de treino: 72%
 - Acurácia de teste: 71%
 - Recall Terapia Adjuvante: 83%
 - Recall Terapia Neoadjuvante: 53%
 - F-1 Score Terapia Adjuvante: 78%
 - F-1 Score Terapia Neoadjuvante: 59%

Servindo de apoio para as métricas, também utilizamos a matriz de confusão para complementar nossas análises e então escolher um modelo candidato. A matriz de confusão nada mais é que uma tabela que mostra como um modelo preditivo está acertando ou errando suas previsões. Na matriz, cada linha representa as previsões feitas pelo modelo para uma classe específica, enquanto cada coluna representa as instâncias reais dessa classe. Estas são as matrizes de precisão que obtivemos para cada um dos 5 algoritmos de predição testados:

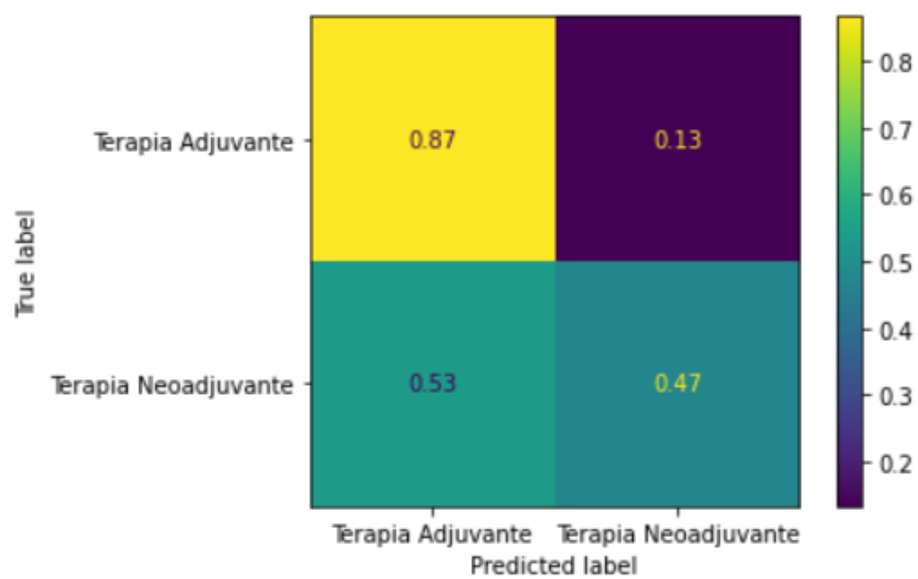
- KNN



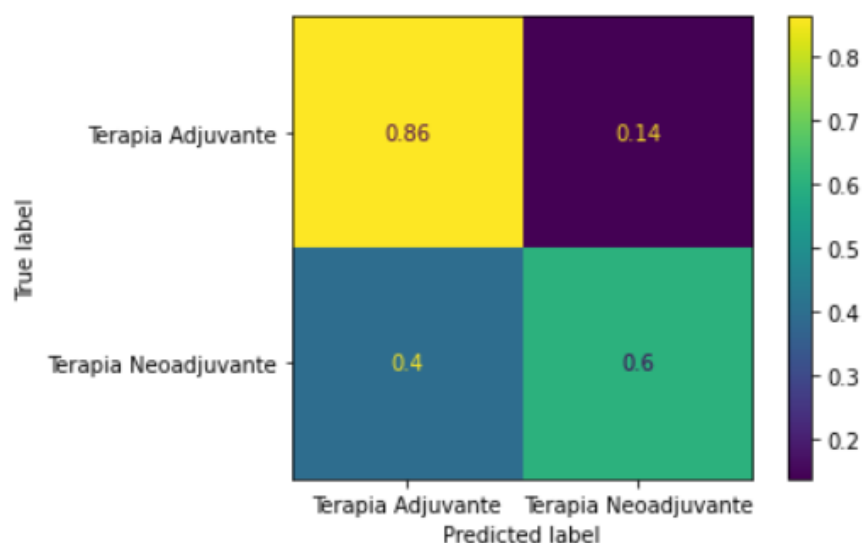
- Random Forest



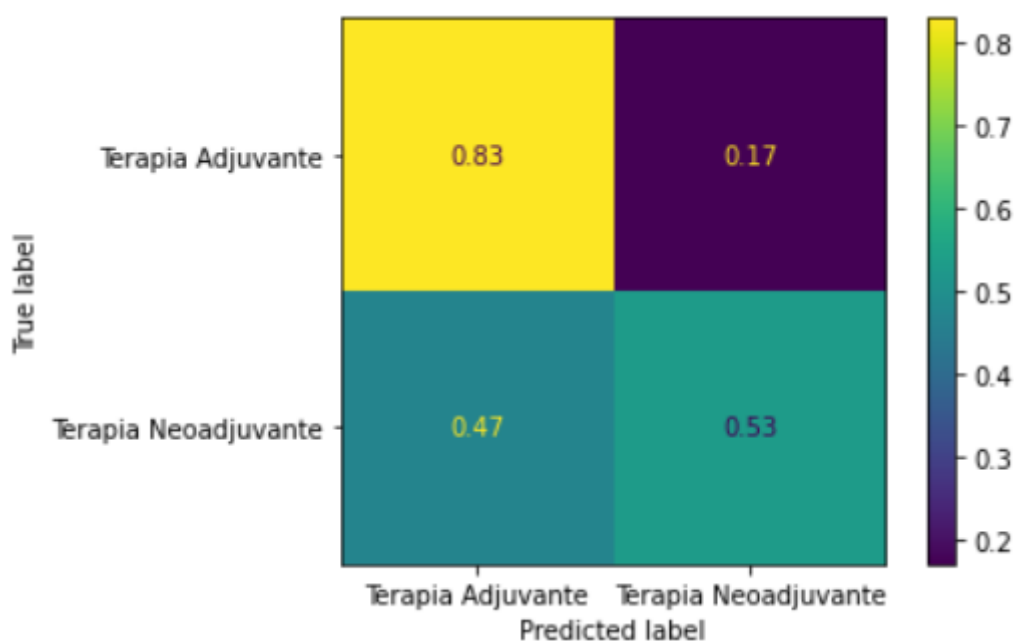
- Naive Bayes



- Decision Tree



Regressão Logística



Escolha do Modelo Candidato

Após avaliar o desempenho de diversos algoritmos de predição, é preciso escolher um modelo candidato para realizar a tarefa de predição em questão. Nesse sentido, analisar as métricas de cada algoritmo pode ajudar a identificar qual é o mais adequado para a tarefa. No nosso caso, optamos pelo **Decision Tree** como modelo candidato, uma vez que ele apresentou uma boa acurácia de teste (cerca de 76%), sem estar em um cenário de overfitting ou underfitting. Além disso, as métricas de Recall e F1-Score foram mais altas nesse algoritmo em comparação com os demais testados. Outro fator que contribuiu para a escolha do Decision Tree foi o fato de que a matriz de confusão apresentou valores mais próximos de 1 para os verdadeiros positivos e verdadeiros negativos.

Embora o Random Forest tenha obtido a maior acurácia de treino e teste entre todos os modelos testados, descartamos sua escolha como modelo candidato, uma vez que ele estava em cenário de overfitting. Portanto, é importante avaliar não apenas a acurácia, mas também outras métricas e aspectos do modelo, como a matriz de confusão e a presença de overfitting ou underfitting, para escolher o modelo mais adequado para a tarefa de predição.

4.4. Comparação de Modelos

a) Escolha da métrica e justificativa.

Durante a modelagem, foi optado seguir a abordagem de retirar os casos considerados como fracasso, para que o modelo classifique, de acordo com os dados do paciente, qual tratamento será mais indicado para que ele tenha sucesso.

Os modelos em si possuem diversas métricas, como:

1. Acurácia: proporção de previsões corretas em relação ao número total de previsões.
2. Precisão: proporção de previsões positivas corretas em relação a todas as previsões.
3. Recall: proporção de previsões positivas corretas em relação a todas instâncias verdadeiramente positivas.
4. F1 Score: média harmônica entre precisão e recall.

Nesse contexto, para o projeto, a métrica mais importante é o F1 Score. Esta métrica é extremamente importante e útil quando as classes não estão distribuídas igualmente. Nesse sentido, as métricas que ignoram o desbalanceamento, como por exemplo a acurácia, podem ser enganosas porque não levam em consideração a proporção entre as classes, fazendo com que o modelo dê preferência para prever a classe majoritária, pois assim aumentará a própria acurácia. Para avaliar quando as classes não estão distribuídas igualmente, é necessário que a diferença seja maior que 60/40, ou seja, 60% pertencer a uma classe e 40% pertencer a outra classe. No caso do projeto, a divisão entre as classes está em: cerca de 60% para Terapia Adjuvante e em torno de 40% para Terapia Neoadjuvante.

Além disso, o F1 Score é interessante pois há uma necessidade de equilibrar tanto o recall quanto a precisão. Uma alta precisão garante que o modelo não classifique falsamente a Terapia Neoadjuvante (nossa instância negativa) como Terapia Adjuvante (nossa instância positiva), enquanto um alto recall garante que o modelo detecte a maioria da instância positiva. Dessa forma, a métrica principal escolhida para o projeto é o F1 Score.

b) Modelos otimizados.

Modelos otimizados são modelos com hiperparâmetros ajustados, com intuito de melhorar sua performance. A otimização do modelo pode ser feita encontrando a melhor combinação de valores de hiperparâmetros, sendo fundamental na construção de um modelo preditivo, pois consegue aumentar significativamente a eficiência e a precisão, consequentemente melhorando o desempenho do modelo como um todo. Além disso, ajuda também a evitar problemas como overfitting, que é quando o modelo é ajustado demais aos dados de treinamento (o modelo decora o comportamento dos dados de treinamento) e não consegue generalizar para novos dados, resultando em um score muito alto no treinamento e um score baixo nos dados de teste.

Nesse contexto, existem várias ferramentas que auxiliam no ajuste dos hiperparâmetros dos modelos, como por exemplo:

1. GridSearchCV: executa uma busca exaustiva em uma grade pré-definida de hiperparâmetros e retorna a melhor combinação de parâmetros.
2. RandomizedSearchCV: executa uma busca aleatória em um espaço pré-definido de hiperparâmetros e retorna a melhor combinação de parâmetros.
3. Cross_val_score: fornece uma pontuação de validação cruzada para um modelo com determinados hiperparâmetros.
4. Learning_curve: plota a curva de aprendizado de um modelo para ajudar a determinar se ele está sofrendo de overfitting ou underfitting.
5. Validation_curve: plota uma curva de validação cruzada para um determinado hiperparâmetro para ajudar a determinar o valor ótimo desse hiperparâmetro

Dessa forma, foi optado por utilizar no projeto a ferramenta 'GridSearch', pois ela passa por todas as combinações de parâmetros possíveis, retornando a melhor.

Em relação aos modelos, foi testado o ajuste em 5 diferentes:

1. KNN (K-NEarest Neighbors)
2. Decision Tree
3. Regressão Logística
4. Random Forest
5. Naive Bayes

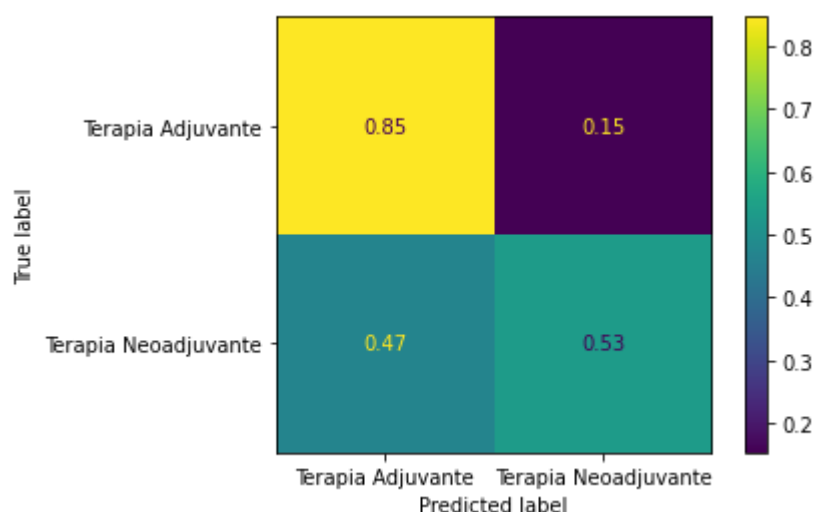
KNN (K-NEarest Neighbors)

O K-Nearest Neighbors (KNN) é um algoritmo de aprendizado de máquina simples e popular que pode ser utilizado para resolver problemas de classificação. A ideia básica do KNN é encontrar os "vizinhos" mais próximos de um novo exemplo de dados e classificá-lo com base nas classes desses vizinhos.

Para entender melhor como o KNN funciona, imagine que temos um conjunto de dados com vários exemplos de flores e suas características, como comprimento e largura das pétalas e sépalas. Cada flor pertence a uma das três classes: rosa, lírio e margarida. Agora, suponha que queremos classificar uma nova flor com base em suas características.

O KNN faria isso encontrando os exemplos de flores mais próximas da nova flor em termos de suas características. Isso é feito medindo a distância entre os exemplos de treinamento e a nova flor. O KNN então classifica a nova flor com base nas classes dos exemplos de treinamento mais próximos. Por exemplo, se os três exemplos de treinamento mais próximos da nova flor forem rosas, o KNN classificaria a nova flor como uma rosa.

Para o modelo em questão, sem a otimização o f1_score é de 78.68%. Nesse sentido, trazendo a matriz de confusão, temos o seguinte resultado:



Para a otimização do modelo usando GridSearch, foi utilizado os seguintes parâmetros:

- 'n_neighbors': [3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47],
- 'weights': ['uniform', 'kd_tree', 'brute'],
- 'metric': ['minkowski', 'euclidean', 'manhattan'],

Dessa forma, os parâmetros encontrados que maximizam o score são:

- 'metric': 'manhattan',
- 'n_neighbors': 39,
- 'weights': 'uniform'

O f1_score médio utilizando os parâmetros ótimos é 71,5%. Isso ocorre porque a busca por hiperparâmetros pode aumentar a complexidade do modelo, levando a overfitting e baixo desempenho na avaliação, ou seja, nesse caso é interessante não fazer a otimização de hiperparâmetros para não aumentar a complexidade do modelo.

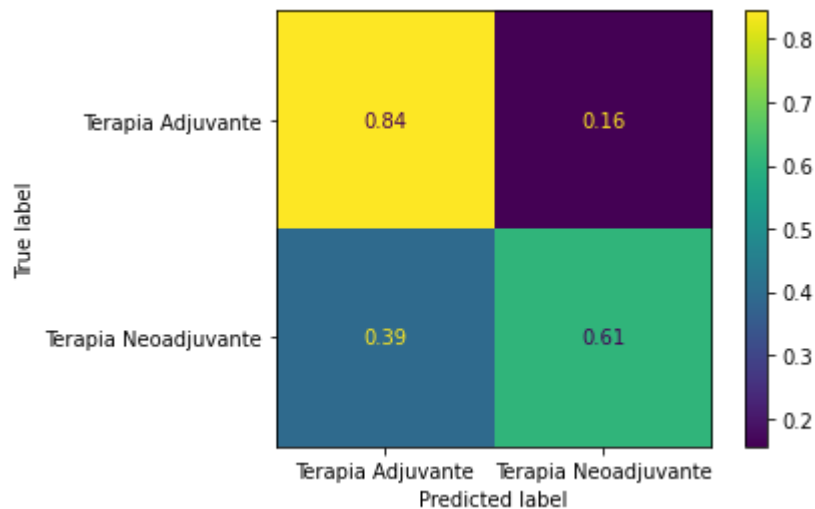
Decision Tree

A Árvore de Decisão é um algoritmo de aprendizado de máquina que cria uma "árvore" de perguntas para decidir qual é a melhor resposta.

Por exemplo, se quisermos decidir se uma fruta é uma maçã ou uma laranja, podemos começar perguntando se ela é vermelha ou laranja. Se for vermelha, a resposta é uma maçã. Se for laranja, precisamos fazer mais perguntas, como se ela tem um talo ou não.

A Árvore de Decisão é útil porque nos ajuda a tomar decisões baseadas em dados. Ele pode ser usado para prever coisas como se um cliente vai comprar um produto ou não, ou se um paciente tem uma doença ou não.

Para o modelo em questão, sem a otimização o f1_score é 79%. Nesse sentido, trazendo a matriz de confusão, temos o seguinte resultado:



Para a otimização do modelo usando GridSearch, foi utilizado os seguintes parâmetros:

- 'criterion': ['gini', 'entropy']
- 'max_depth': [None, 1, 2, 3, 5, 7, 9],
- 'max_features': ['sqrt', 'log2'],
- 'min_samples_leaf': [1, 2, 4],
- 'min_samples_split': [2, 5, 10],

Dessa forma, os parâmetros encontrados que maximizam o score são:

- 'criterion': 'gini',
- 'max_depth': None,
- 'max_features': 'sqrt',
- 'min_samples_leaf': 1,
- 'min_samples_split': 2

O f1_score utilizando os parâmetros ótimos é 70,33%. Dessa forma, o comportamento é semelhante ao primeiro modelo, onde a busca por hiperparâmetros pode aumentar a complexidade do modelo, levando a overfitting e baixo desempenho na avaliação, ou seja, nesse caso é interessante não fazer a otimização de hiperparâmetros para não aumentar a complexidade do modelo.

Regressão Logística

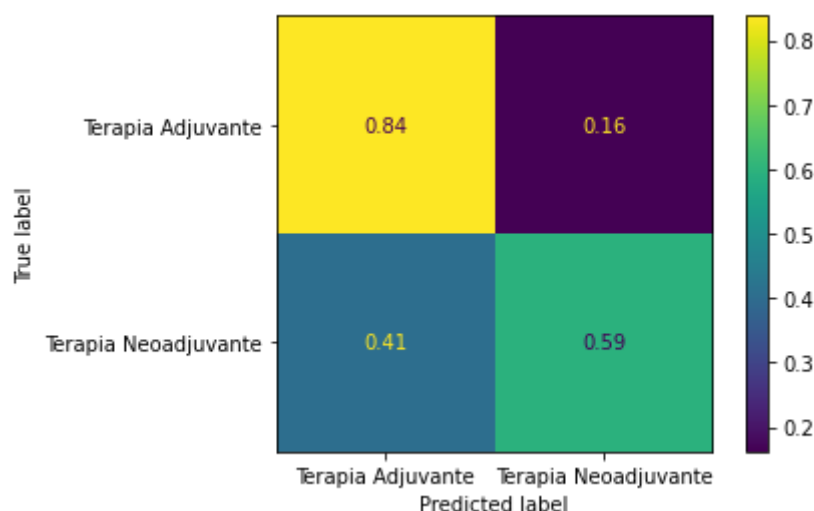
A Regressão Logística é um algoritmo de aprendizado de máquina usado para prever a probabilidade de um evento acontecer, como a probabilidade de um paciente ter uma doença com base em informações como idade, sexo e histórico médico.

O algoritmo usa uma função matemática chamada função logística para transformar a probabilidade calculada em uma escala de 0 a 1, onde valores próximos de 0 indicam baixa probabilidade e valores próximos de 1 indicam alta probabilidade.

A Regressão Logística é útil porque permite que você avalie como diferentes fatores podem afetar a probabilidade de um evento acontecer. Por exemplo, se você está tentando

prever se um cliente irá cancelar uma assinatura, você pode considerar fatores como idade, gênero, histórico de pagamentos e outras informações para avaliar a probabilidade de cancelamento.

Para o modelo em questão, sem a otimização, o f1_score médio é de 78,62%. Nesse sentido, trazendo a matriz de confusão, temos o seguinte resultado:



Para a otimização do modelo usando GridSearch, foi utilizado os seguintes parâmetros:

- 'penalty': ['None', 'l1', 'l2', 'elasticnet'],
- 'solver': ['lbfgs', 'newton-cg', 'liblinear', 'sag', 'saga'],
- 'max_iter': [100, 1000, 2500, 5000]

Dessa forma, os parâmetros encontrados que maximizam o score são:

- 'max_iter': 2500,
- 'penalty': 'l1',
- 'solver': 'saga'

O f1_score médio utilizando os parâmetros ótimos é 78,55%.

Random Forest

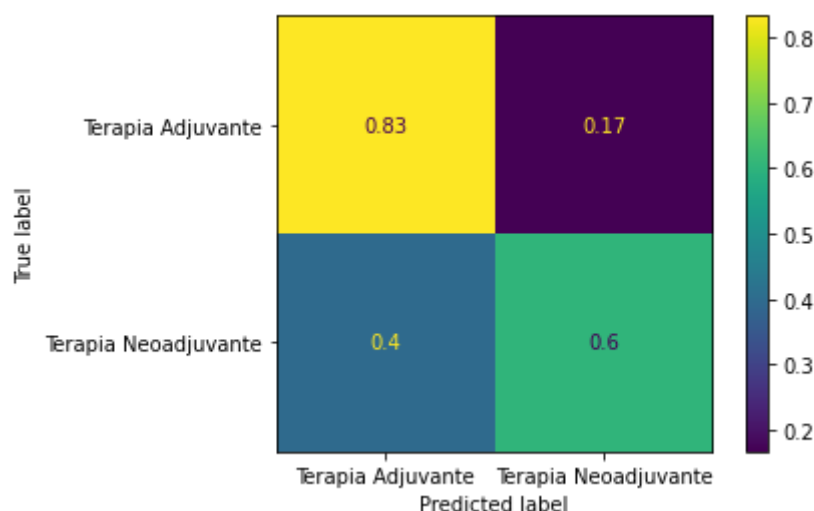
O Random Forest é um algoritmo de aprendizado de máquina que pode ser usado tanto para classificação quanto para regressão. A ideia principal do Random Forest é construir várias árvores de decisão e combiná-las para obter uma predição mais precisa.

Árvores de decisão são estruturas em forma de árvore que ajudam a tomar decisões com base em uma série de regras simples. Cada nó da árvore representa uma condição, como "se a idade é maior que 30 anos" ou "se a renda é menor que 50 mil". As decisões são tomadas seguindo os ramos da árvore até chegar a uma folha, que representa a classificação final.

Quando uma nova amostra é apresentada ao modelo Random Forest, as árvores individuais dão suas previsões e a previsão final é a que tem mais votos. Por exemplo, se cinco

árvores preveem que uma determinada amostra é da classe A e quatro preveem que é da classe B, o modelo Random Forest classifica a amostra como da classe A.

Para o Random Forest, sem a otimização o $f1_score$ é cerca de 78,6%. Nesse sentido, trazendo a matriz de confusão, temos o seguinte resultado:



Para a otimização do modelo usando GridSearch, foi utilizado os seguintes parâmetros:

- "bootstrap": [True],
- "max_depth": [6, 7, 8, 9, 10, 12, 14, 16, 18],
- "max_features": ['auto', 'sqrt'],
- "min_samples_leaf": [3, 4, 5, 6, 7, 8],
- "min_samples_split": [2, 3, 4, 5, 6, 7, 8],
- "n_estimators": [100, 350]

Dessa forma, utilizando o GridSearch, os parâmetros encontrados que maximizam o score são:

- 'bootstrap': True,
- 'max_depth': 6,
- 'max_features': 'auto',
- 'min_samples_leaf': 3,
- 'min_samples_split': 2,
- 'n_estimators': 100

O $f1_score$ médio utilizando os parâmetros ótimos é: 78,2%

Naive Bayes

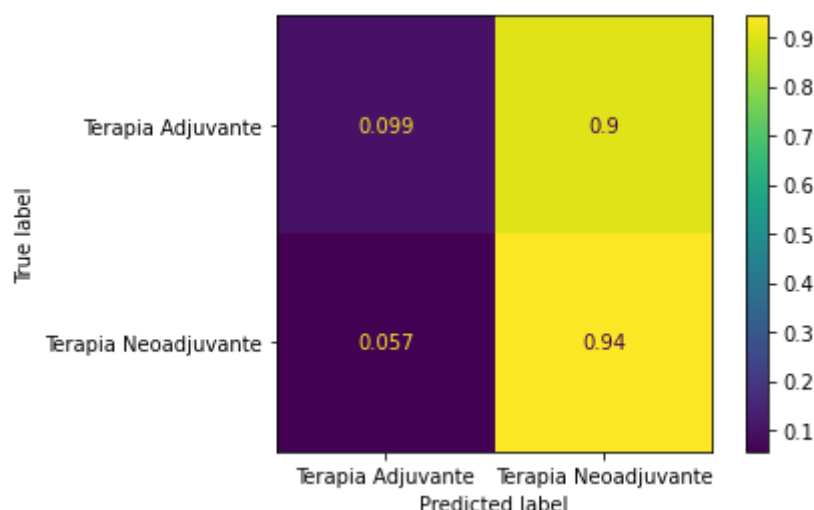
O Naive Bayes é um algoritmo de aprendizado de máquina que pode ser usado para classificação de dados em diferentes categorias. Ele é baseado no teorema de Bayes, que

afirma que a probabilidade de uma hipótese (como uma categoria) ser verdadeira é proporcional à probabilidade das evidências (como as características dos dados) sob essa hipótese.

A ideia principal do Naive Bayes é calcular a probabilidade de um ponto de dados pertencer a uma determinada categoria com base na probabilidade das características do ponto de dados em cada categoria. O algoritmo assume que as características são independentes entre si, ou seja, a presença de uma característica não afeta a probabilidade de outra característica estar presente.

Por exemplo, se tivermos um conjunto de dados de e-mails e quisermos classificar se eles são spam ou não spam, o Naive Bayes calcularia a probabilidade de um e-mail ser spam com base na probabilidade de cada palavra aparecer em e-mails de spam e na probabilidade de cada palavra aparecer em e-mails não spam. Se a probabilidade de um e-mail ser spam for maior do que a probabilidade de não ser spam, o Naive Bayes classificaria o e-mail como spam.

Para o Naive Bayes, sem a otimização o f1_score é cerca de 17,4%. Nesse sentido, trazendo a matriz de confusão, temos o seguinte resultado:



Para a otimização do modelo usando GridSearch, foi utilizado o seguinte parâmetro:

- "var_smoothing": [1e-9, 1e-8, 1e-7, 1e-6, 1e-5]

Dessa forma, utilizando o GridSearch, o argumento encontrado que maximiza o score é:

- 'var_smoothing': 1e-09

O f1_score médio utilizando os parâmetros ótimos é: 17,4%

Definição do modelo escolhido e justificativa:

Ao selecionar o modelo ficamos entre dois algoritmos:

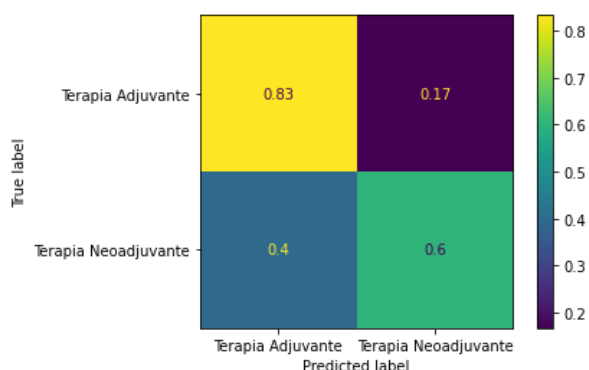
- Random Forest
- Decision Tree

Para decidirmos o selecionado, analisamos alguns pontos, sendo o principal o F1 Score, mas levamos em consideração também a acurácia, recall, e a análise das matrizes de confusão. A análise dessas métricas nos levou ao consenso que o algoritmo de Decision Tree é o mais adequado, já que ele atingia métricas melhores.

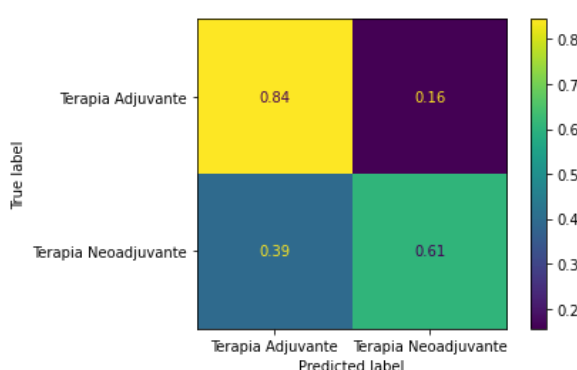
Primeiramente em relação a acurácia: No que tange ao Decision Tree, obtivemos um valor de 0.74 tanto no conjunto de treino quanto de teste, em que, diferente do algoritmo de Random Forest que obtivemos 0.73 no conjunto de teste e 1.0 no conjunto de treino, se tornou evidente que existe uma diferença gigantesca entre os dois conjuntos, fato que leva a acreditar que o algoritmo apresenta overfitting, enquanto o algoritmo de Decision Tree se manteve condizente nos conjuntos de treino e teste.

Em relação ao recall: é possível identificar tanto no modelo de Random Forest quanto no Decision Tree que os algoritmos tinham muito mais dificuldade em identificar os casos verdadeiros positivos de terapia neoadjuvante em relação aos casos verdadeiros positivos de terapia adjuvante. Esse fator é fortemente reforçado quando analisamos as matrizes de confusão de ambos os algoritmos:

Random Forest:



Decision Tree:



Pode-se observar que o algoritmo de Decision Tree consegue afirmar com maior certeza casos verdadeiros positivos de terapia neoadjuvante.

Finalmente, o principal, em relação ao F1 Score: é indiscutível que essa é a métrica mais importante para a decisão do modelo utilizado. Entretanto, Random Forest e Decision Tree apresentaram f1 scores muito próximos, sendo 78.6% e 79%, respectivamente.

Claramente é de suma importância que se faça uma análise para a seleção do algoritmo utilizado. Portanto, o fato do algoritmo de random forest apresentar overfitting, f1 score menor e uma confusão maior nos verdadeiros positivos da terapia neoadjuvante fez com que escolhêssemos o algoritmo de Decision Tree.

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.

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.