# PROJETO AGATHA
# FACULDADE DE MEDICINA DA USP

inteli
instituto
de tecnologia
e liderança

# Controle do Documento

## Histórico de revisões

| Data | Autor | Versão | Resumo da atividade |
|---|---|---|---|
| 30/01/2023 | Yago Araújo | 1.1 | Criação do Documento e personalização conforme o grupo 3 |
| 20/02/2023 | Enya Oliveira Arruda | 1.2 | Preenchimento da seção 1.0 |
| 20/02/2023 | Enya Oliveira Arruda | 1.3 | Preenchimento da seção 2.0 |
| 16/03/2023 | Isabela Rocha | 1.4 | Preenchimento da seção 3.0 |
| 21/02/2023 | Yago Araújo<br>Luis Miranda | 1.5 | Atualização da seção 4.1.4 |
| 21/02/2023 | Luis Miranda | 1.6 | Atualização da seção 4.1.1 |
| 04/02/2023 | Marcelo Maia | 1.7 | Preenchimento da seção 4.1.1.2 |
| 21/02/2023 | Luis Miranda | 1.8 | Atualização da seção, 4.1.2 |
| 05/02/2023 | Luis Miranda | 1.9 | Preenchimento da seção 4.1.5 |
| 22/02/2023 | Isabela Rocha<br>Marcelo Maia<br>Luis Miranda | 2.0 | Atualização da seção 4.1.6 |
| 21/02/2023 | Enya Oliveira Arruda<br>Thomaz Barboza<br>Luis Miranda | 2.1 | Atualização da seção 4.1.7 |
| 06/02/2023 | Enya Oliveira Arruda | 2.2 | Preenchimento da seção, 4.1.3 |
| 23/02/2023 | Luis Miranda | 2.3 | Preenchimento da seção 4.2 |
| 26/02/2023 | Enya Oliveira<br>Luis Miranda | 2.4 | Preenchimento da seção 4.2.1; 4.2.2; 4.2.3 |
| 26/02/2023 | Enya Oliveira Arruda | 2.5 | Preenchimento da seção 4.2.4 - política de privacidade |
| 07/03/2023 | Yago Araújo | 2.6 | Preenchimento da seção 4.3 - Métricas |
| 11/03/2023 | Fabio Piemonte | 2.7 | Preenchimento da seção 4.3 - a) e c) |

| | | | |
|---|---|---|---|
| 22/03/2023 | Marcelo Maia | 4.4 | Preenchimento da seção 4.4 b) |
| 23/03/2023 | Thomaz Klifson | 4.5 | Finalizando o preenchimento da seção 4.4 b) |
| 27/03/2023 | Enya Oliveira | 4.6 | Correção Ortográfica e Gramatical |

# Sumário

# 1. Introdução

O Hospital das Clínicas de São Paulo é uma instituição pública de saúde ligada à Universidade de São Paulo (USP). Fundado em 1944, é considerado o maior complexo hospitalar da América Latina e um dos mais importantes do país. O hospital possui 11 unidades especializadas em diversas áreas, como cardiologia, oncologia, neurologia, pediatria e transplantes, entre outras. Além disso, conta com um centro de diagnóstico e tratamento avançado, com equipamentos de última geração para exames e cirurgias.

O Hospital das Clínicas é referência em atendimento de alta complexidade, recebendo pacientes de todo o Brasil e até de outros países. Além disso, é um importante centro de pesquisa e formação de profissionais da área de saúde, com programas de pós-graduação e residência médica reconhecidos nacional e internacionalmente. Apesar dos desafios enfrentados pelo sistema de saúde público no Brasil, o Hospital das Clínicas de São Paulo se destaca por sua qualidade e eficiência, com uma equipe de profissionais altamente qualificados e uma estrutura moderna e bem equipada.

# 2. Objetivos e Justificativa

## 2.1. Objetivos

O objetivo do projeto é descobrir um padrão preditivo entre pacientes diagnosticados com câncer, a fim de determinar para cada caso o tipo mais adequado de terapia de tratamento, seja ela neoadjuvante ou adjuvante. O tratamento para o câncer de mama é um dos mais procurados no hospital das clínicas e devido à ampla variedade de variáveis ao designar o tratamento mais viável para as pacientes, iremos determinar uma ferramenta que auxilie na prescrição do diagnóstico.

## 2.2. Proposta de Solução

Estamos desenvolvendo um modelo preditivo proposto para definir um tratamento de câncer baseado em dados clínicos e de saúde do paciente, incluindo informações sobre o tipo e estágio do câncer, idade, histórico médico e outros fatores relevantes.

O objetivo é fornecer uma recomendação de tratamento personalizada e baseada em evidências para cada paciente, que leve em consideração todos os fatores relevantes e ajude a equipe médica a tomar uma decisão informada sobre o melhor curso de ação. Existem duas opções de tratamento, a coadjuvante e a neoadjuvante, e a ferramenta auxiliará o oncologista e sua equipe na precisão da prescrição.

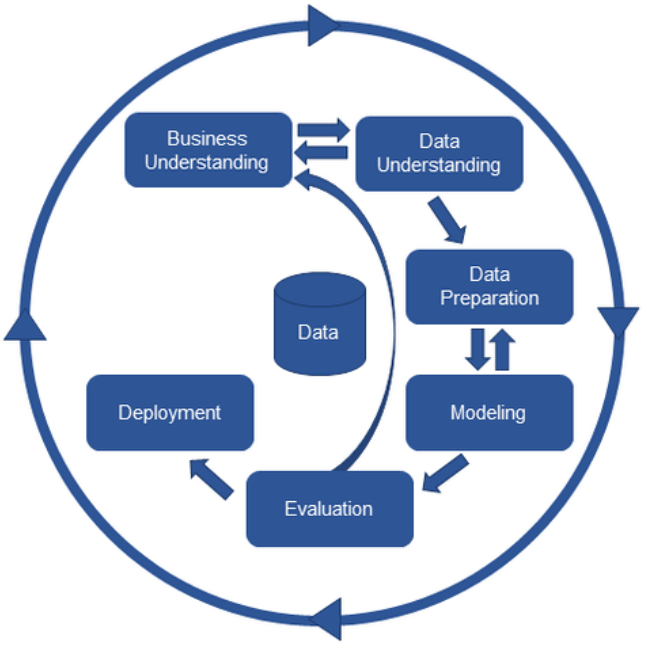
## 2.3. Justificativa

O câncer de mama é uma doença complexa e pode afetar diferentes pessoas de maneiras diferentes. Um modelo preditivo permite uma abordagem mais personalizada ao tratamento, levando em consideração as características individuais do paciente. O modelo preditivo é alimentado por dados clínicos e de saúde dos pacientes, como por informações sobre as últimas descobertas científicas e as melhores práticas médicas disponíveis para o tratamento do câncer. Isso significa que a recomendação de tratamento é baseada em evidências sólidas e atualizadas, ajudando a minimizar o risco de erro humano na definição do tratamento. Além disso, o modelo permite a integração de uma grande quantidade de dados e informações, o que é difícil de ser feito de forma manual.

# 3. Metodologia

## 3.1 CRISP-DM

A metodologia CRISP-DM é um processo iterativo e ágil, que consiste em diversas fases relacionadas ao tratamento de dados. Essa metodologia permite retornar às etapas anteriores para realizar melhorias, adaptações e incrementos. Em nosso projeto, optamos por utilizar metodologias ágeis, como o CRISP-DM, para garantir um melhor relacionamento com o parceiro e transformar o grande volume de dados em informações úteis para a gestão e tomada de decisões da equipe. Durante o projeto, essa metodologia foi amplamente utilizada devido aos incrementos e mudanças contínuas que ocorreram em cada *sprint* e validação com nosso parceiro.




**Anexo 1 - CRISP-DM.**

Descrevendo as etapas do CRISP-DM, temos:

1. **Business Understanding**

   A metodologia começa com a identificação das necessidades do projeto, o que inclui identificar o tema a ser elaborado, estudar e compreender o processo, compreender o negócio e assim por diante. Ao concluir esta etapa, espera-se que sejam respondidas

perguntas importantes como: qual é o problema a ser resolvido, por que ele é relevante, entre outras.

2. **Data Understanding**
   Nesta etapa, o objetivo é coletar, organizar e documentar todos os dados disponíveis para a análise exploratória. Caso seja necessário, será preciso utilizar várias fontes de dados e considerar como e quando integrá-las. É importante arquitetar e estudar a melhor maneira de extrair informações, levando em conta a origem de cada conteúdo e os softwares que serão utilizados. Os dados também devem ser revisados antes da extração. O processo de extração deve ser definido em conjunto com toda a equipe, identificando as áreas e pessoas que serão envolvidas e mapeadas para função. Além disso, é fundamental definir os formatos de arquivos e variáveis com antecedência para evitar inconsistências.

3. **Data Preparation**
   Nesta etapa, é necessário realizar o tratamento dos dados para garantir a coerência das informações. É preciso resolver inconsistências, erros e ausência de valores para que possamos selecionar amostras aleatórias e utilizá-las para treinamento e testes. Além disso, é importante definir métodos de avaliação dos modelos para que possamos avaliar o desempenho dos experimentos antes mesmo de apresentá-los à equipe de negócios. Durante essa etapa, é possível tratar valores nulos ou inconsistentes, criar novos dados (se necessário), selecionar atributos relevantes para o modelo e escolher métodos adequados para avaliação do modelo e seleção de amostras.

4. **Modeling**
   Nesta fase, estabelecemos os modelos que serão testados e justificamos as razões por trás de cada escolha. Realizamos avaliações do modelo, estimamos hiperparâmetros e etc.

5. **Evaluating**
   Nesta etapa, é necessário realizar uma avaliação dos resultados e explorar todas as possíveis variações que os dados possam apresentar. Isso envolve a análise de quaisquer fatores que possam ter sido negligenciados e a determinação em que medida o modelo atende aos objetivos do negócio. Se o modelo não estiver apresentando o desempenho esperado, é preciso voltar à primeira etapa para compreender melhor o negócio e os dados.

6. **Deployment**
   Nessa fase, o modelo é implantado em um protótipo e enviado para produção. É necessário monitorar o desempenho do modelo durante um período previamente acordado com a equipe de negócios e realizar uma estratégia de rollout.

## 3.2 Ferramentas

As seguintes plataformas foram empregadas como ferramentas e bibliotecas para aprimorar o uso do CRISP-DM:

| | | |
|---|---|---|
|  Google Colab |  Google Drive |  GitHub |

**Anexo 2 - Ferramentas**

Utilizamos o Colab por ser uma plataforma de desenvolvimento integrado baseado em nuvem que permite aos integrantes escrever, executar e compartilhar códigos em Python. É especialmente útil para projetos de machine learning, pois permite a utilização de GPUs e TPUs gratuitamente. Além disso, é possível conectar o Colab ao Google Drive, permitindo o acesso de todos os integrantes a arquivos e pastas armazenados na nuvem.

Já o Google Drive é um serviço de armazenamento em nuvem que permite armazenar, compartilhar e acessar arquivos de qualquer lugar. Com a integração do Google Drive ao Colab, é possível compartilhar arquivos e pastas entre os membros da equipe.

O GitHub é uma plataforma de hospedagem de código-fonte que permite que os membros da equipe trabalhem juntos no projeto. Com o GitHub, é possível realizar o controle de versão do código-fonte, colaborar com os outros integrantes da equipe e garantir a integridade do código.

Ao utilizar o Colab, Google Drive e GitHub juntos, os membros da equipe podem trabalhar juntos em tempo real e de forma colaborativa. Essas ferramentas tornam o processo de desenvolvimento muito mais eficiente e produtivo, além de garantir a integridade do projeto e a rastreabilidade de alterações e melhorias.

## 3.3 Técnicas empregadas

Para aprimorar a nossa abordagem e obter resultados mais precisos e eficazes na avaliação dos modelos treinados, optamos por utilizar diversas bibliotecas e técnicas em nosso projeto de modelo preditivo.

Entre as técnicas utilizadas, destacamos o uso de bibliotecas do 'scikit-learn', como a 'accuracy_score' e 'precision_score', que nos permitiram avaliar os modelos com maior precisão. Além disso, empregamos as bibliotecas' pandas' e 'numpy' para tratamentos de dados, incluindo técnicas como 'label encoding'.

Para apresentar os resultados do projeto de forma visual e clara, utilizamos bibliotecas como 'Seaborn' para criar boxplots e gráficos de colunas, a 'confusion_matrix' do 'scikit-learn' e o matplotlib para gerar gráficos com os dados obtidos. Essas técnicas e ferramentas foram fundamentais para aprimorar nosso modelo preditivo e obter resultados mais precisos e confiáveis.

# 4. Desenvolvimento e Resultados

## 4.1. Compreensão do Problema

### 4.1.1. Contexto da indústria

O Hospital das Clínicas de São Paulo (HCSP) é uma instituição de destaque na área da saúde no Brasil. Fundado em 1940, este complexo hospitalar universitário localizado na cidade de São Paulo é considerado um dos principais centros médicos da América Latina, com uma ampla gama de serviços de saúde prestados.

Com uma infraestrutura moderna e tecnológica, o HCSP se destaca como uma instituição de referência em diversas áreas médicas, com foco para cardiologia, oncologia e neurologia. Além de prestar atendimento de qualidade com a saúde pública, o hospital também é responsável pela formação de novos profissionais da saúde, como médicos e enfermeiros.

Sua administração está sob a responsabilidade da Universidade de São Paulo (USP), uma das maiores e mais prestigiadas instituições de ensino superior do Brasil e da América Latina, garantindo assim a qualidade e excelência nos serviços prestados aos pacientes. O hospital também conta com colaborações com diversas empresas e outras instituições de ensino, como o Inteli, o que possibilita a condução de pesquisas e o avanço de novas tecnologias para a área da saúde.

Falando em concorrentes, é importante destacar que o HCSP possui uma missão específica como hospital público, o que o diferencia dos hospitais privados. No entanto, no setor público, há outros hospitais universitários e de ensino que podem ser vistos como concorrentes, como o Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG), o Hospital Universitário de Brasília (HUB) e o Hospital Universitário da Universidade Federal do Rio de Janeiro (UFRJ). Adicionalmente, no setor privado, existem vários hospitais de São Paulo que competem em termos de qualidade de atendimento e serviços oferecidos.

Em relação às tendências, pode-se afirmar que o HCSP está aderindo às tendências comuns na indústria da saúde, como a crescente utilização de tecnologia, o enfoque na medicina personalizada e a busca por maior eficiência no atendimento. Ademais, como um hospital público, é provável que o HC esteja enfrentando desafios relacionados à escassez de financiamento e limitação de recursos.

No cenário em geral, a indústria de saúde no Brasil tem experimentado um crescimento constante ao longo dos anos, e a presença do HCSP é fundamental para fortalecer esse setor. Com uma equipe altamente qualificada e dedicada à saúde da população, o hospital é visto

como referência por outras instituições e contribui para melhorar a qualidade de vida da população.

Em síntese, a indústria do HCSP é caracterizada pela excelência na prestação de serviços da saúde pública, pela formação de novos profissionais, pela realização de pesquisas e pelo avanço de novas tecnologias. Tudo isso confere ao hospital uma posição de destaque na indústria da saúde do Brasil.

### 4.1.1.2 As 5 forças de porter

**-Rivalidade entre concorrentes:**

Considerando que o ICESP(Instituto do Câncer do Estado de São Paulo) é uma instituição estatal e sem fins lucrativos, não possui concorrentes diretos. Contudo, é necessário atentar-se aos métodos e tratamentos oferecidos pelos principais hospitais privados especializados em tratamento de câncer no Brasil, para que assim o instituto mantenha-se atualizado e sendo um centro de tratamento com muita qualidade e seja capaz de oferecer serviços equiparáveis ao setor privado.

**-Poder de barganha dos Fornecedores:**

Novamente, por ser uma instituição governamental o capital disponível é limitado de acordo com quanto o Estado irá destinar para a área da saúde. Além disso, tendo em vista que os produtos necessários para a manutenção e avanço do instituto são altamente especializados e de alto valor, assim como são indispensáveis para o funcionamento do ICESP. Dessa forma, os fornecedores possuem um alto poder de barganha , principalmente devido ao seus bens fornecidos serem de um nicho muito específico e de alto valor, mas também por não ser um mercado tão diluído.

**-Poder de barganha dos compradores:**

Em relação ao ICESP, o poder de barganha é baixo, pois os clientes/pacientes estão utilizando-se de um serviço público, que é prestado de maneira gratuita. Logo, uma possível barganha de preço não se a situação.

Por outro lado, considerando o poder de barganha dos clientes em relação ao software desenvolvido é altíssimo .

**-Ameaça de novos entrantes:**

Com o avanço cada vez maior das instituições de pesquisa, o risco da ameaça de novos entrantes é alto, pois a iniciativa privada tem mais capital. Além disso, o risco do surgimento de instituições privadas com soluções mais adequadas e viáveis é alto e deve ser levado em consideração pelo ICESP.

**-Ameaça de produtos substitutos:**

O crescimento exponencial do uso de IA na área da saúde representa um alto risco para o modelo preditivo desenvolvido, tendo em vista que já existem modelos que atualmente são utilizados para a prescrição de tratamento para outras doenças , mas poderiam ser convertidos para a utilização para o câncer de mama.

Além disso, o desenvolvimento de modificação de genes baseado na sequenciação do genoma está sendo bastante pesquisado e, caso concretizado, não haveria a necessidade do modelo preditivo desenvolvido.

## 4.1.2. Análise SWOT

Na análise SWOT, buscamos definir os pontos fortes e fracos, oportunidades e ameaças do Hospital das Clínicas, com influência de forças externas (algo que o hospital não possui controle) e internas (que podem ser controladas pelo hospital). Em seguida da matriz, disponibilizamos a legenda de cada ponto levantado


**Anexo 3 - Análise SWOT.**

**Pontos Fortes:**

- **Reputação**: O Hospital das Clínicas possui um amplo reconhecimento por sua excelência em pesquisa e atendimento médico público, fazendo com que possua pacientes de todo o país.
- **Corpo docente**: O hospital possui uma equipe altamente qualificada que trabalha juntos para oferecer o melhor atendimento aos pacientes.
- **Pesquisa**: O hospital é líder em pesquisa médica, tanto no Brasil quanto no mundo.

- **Infraestrutura**: O hospital possui instalações modernas e tecnológicas, incluindo laboratórios, equipamentos médicos e clínicas de última geração.

**Fraquezas:**

- **Financiamento**: Apesar de ter uma equipe altamente qualificada, o hospital pode sofrer com a falta de recursos financeiros para investir em pesquisas, infraestrutura e tecnologia.
- **Lista de espera**: Pelo fato de ter uma equipe altamente qualificada e com serviços gratuitos para a população, o hospital acaba sendo muito procurado, o que pode resultar em listas de espera longas para procedimentos e consultas.

**Oportunidades:**

- **Expansão**: Há potencial para o hospital expandir seu alcance de ensino e serviços por meio de parcerias com outras universidades, instituições e atendimento médico remoto, o que pode melhorar a acessibilidade aos pacientes.
- **Tecnologia**: O hospital pode aproveitar as vantagens de tecnologias novas e emergentes para melhorar o atendimento ao paciente, investindo em inteligência artificial para melhorar a eficiência e a qualidade do atendimento.
- **Internacionalização**: O hospital pode aumentar seu impacto global atraindo estudantes, professores e pesquisadores de todo o mundo.

**Ameaças:**

- **Instabilidade política**: A instabilidade política no Brasil pode afetar o financiamento e a estabilidade do hospital e de seu corpo docente, fazendo com que a qualidade de atendimento caia.
- **Desafios globais de saúde**: O hospital pode enfrentar desafios para responder a crises emergentes de saúde global, como pandemias.
- **Concorrência:** O hospital enfrenta a concorrência de novos hospitais escola de medicina, como o Albert Einstein e a Santa Casa, o que pode dificultar a atração de alunos e a manutenção de sua reputação.

### 4.1.3. Planejamento Geral da Solução

**a) Qual é o problema a ser resolvido?**

A evolução do câncer de mama e sua resposta a tratamentos convencionais é muito variável, portanto, é necessário encontrar padrões e relações entre os tratamentos já realizados em inúmeros pacientes para entender qual o melhor tratamento deve ser indicado para um indivíduo.

**b) Qual a solução proposta (visão de negócios)?**

Um modelo preditivo que encontra padrões e relações entre os dados de inúmeros pacientes durante o tratamento do câncer de mama, levando em consideração dados demográficos, histopatológicos, registros dos tumores, como também peso e altura.

**c) Como a solução proposta deverá ser utilizada?**

A solução poderá ser usada a partir de um simples site, onde o médico poderá colocar os dados do paciente para então receber uma recomendação de tratamento da doença, levando em consideração a melhor opção para o seu paciente.

IDEIA - Fornecer a solução via API, possibilitando a implementação dela em softwares já usados por laboratórios/médicos/hospitais.

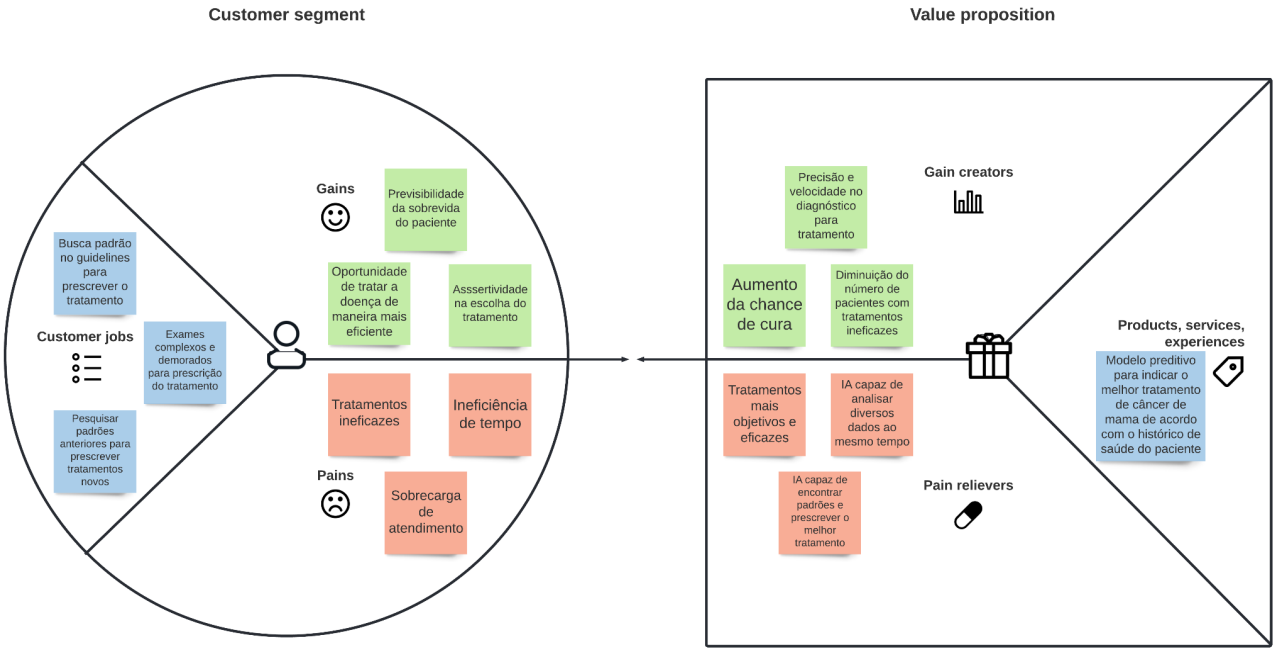
**d) Quais os benefícios trazidos pela solução proposta?**

Auxílio estatístico para a tomada de decisão do tratamento do câncer de mama, levando em consideração os dados do indivíduo e o relacionando com os dados de outros milhares de pacientes, resultando na melhor opção a ser tomada.

**e) Qual será o critério de sucesso e qual medida será utilizada para o avaliar?**

O critério de sucesso para o nosso modelo será baseado nas métricas de avaliação para machine learning relacionando a taxa de sobrevida do indivíduo após o tratamento com o tipo de câncer que ele possui.

## 4.1.4. Value Proposition Canvas




**Anexo 4 - Value Proposition Canvas.**

### Customer segment

Customer jobs:

- Busca padrão no guideline para prescrever o tratamento;
- Exames complexos e demorados para prescrição do tratamento;
- Pesquisar padrões anteriores para prescrever tratamentos novos.

Gains:

- Previsibilidade da sobrevida do paciente;
- Oportunidade de tratar a doença de maneira mais eficiente;
- Assertividade na escolha do tratamento.

Pains:

- Tratamentos ineficazes;
- Ineficiência de tempo;
- Sobrecarga de atendimento.

### Value proposition

Gain creators:

- Precisão e velocidade no diagnóstico para tratamento;

- Aumento da chance de cura;
- Diminuição do número de pacientes com tratamentos ineficazes.

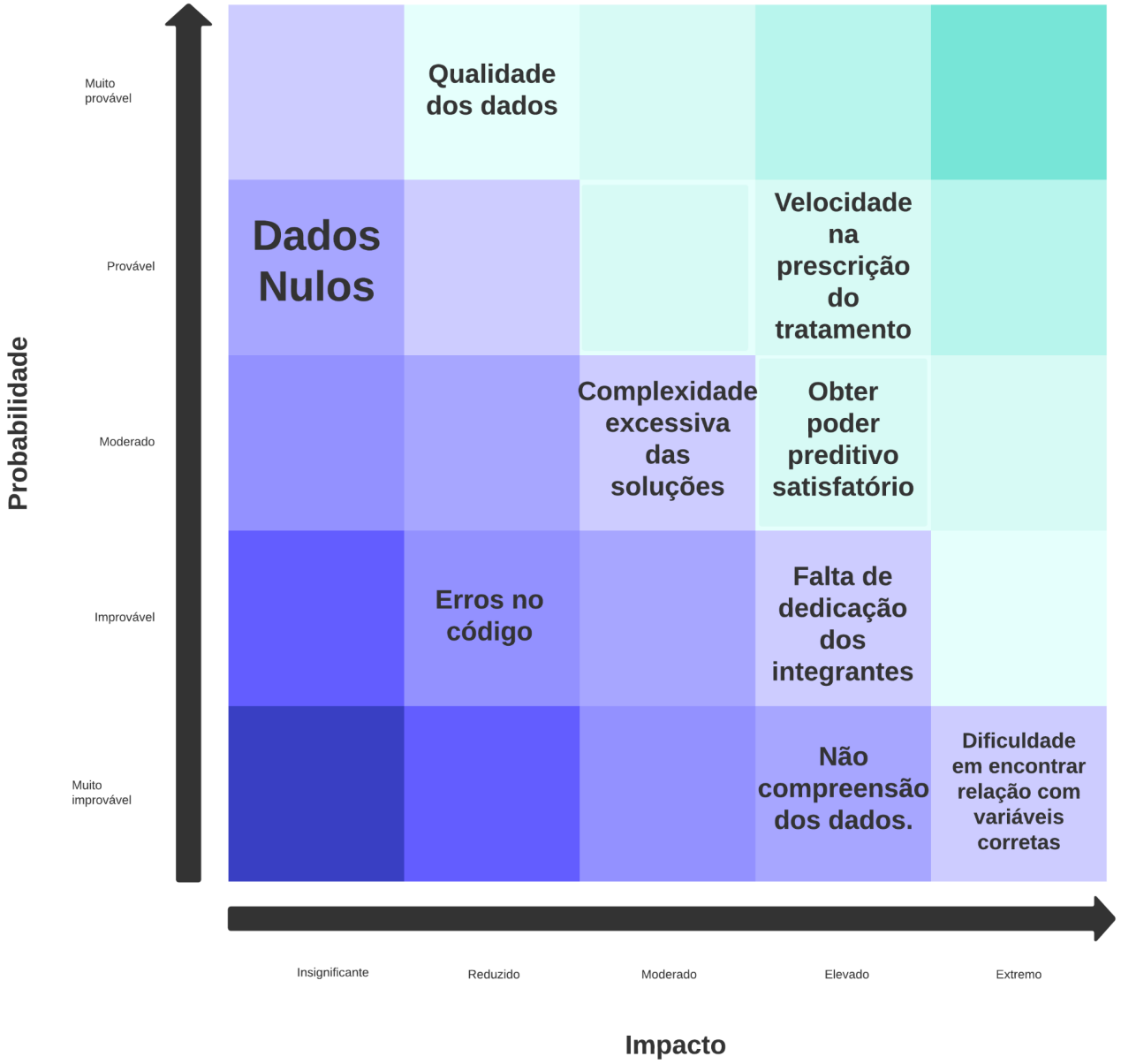
Pain relievers:

- Tratamentos mais objetivos e eficazes;
- IA capaz de analisar diversos dados ao mesmo tempo;
- IA capaz de encontrar padrões e prescrever o melhor tratamento.

Products, services, experiences:

- Modelo preditivo para indicar o melhor tratamento de câncer de mama de acordo com o histórico de saúde do paciente.

## 4.1.5. Matriz de Riscos




Anexo 5 - Matriz de Risco.

## 4.1.6. Personas

- **Persona 1**

Médica Oncologista utilizadora do modelo preditivo para prescrição do tratamento.


**SOBRE**

Marcela ingressou na USP aos 20 anos de idade, sendo sempre muito dedicada e se destacando como uma das melhores alunas de sua turma. Durante o período acadêmico, conheceu seu esposo, também médico, com quem se casou e teve um filho.
Com grande apreço pelo contato interpessoal, Marcela encontra satisfação no atendimento e acompanhamento do tratamento de seus pacientes, buscando sempre fazer a diferença em suas vidas. Foi por isso que optou por se especializar em oncologia, área em que iniciou sua carreira no Instituto do Câncer do Estado de São Paulo (ICESP), buscando assim ter um impacto ainda maior na vida de seus pacientes.

**HABILIDADES**

Matemática

Tecnologia

Tocar piano

**PERSONALIDADE**

Analítica

Extrovertida

Ocupada

Independente

**DESAFIOS**

- Enfrenta um desafio constante em sua prática médica, que é a escolha do melhor tratamento para cada paciente.
- Acredita que a IA pode ser uma ferramenta útil para ajudá-la a tomar decisões mais precisas e individualizadas sobre o tratamento do câncer de mama

**METAS**

- Viajar com a família
- Tratar e curar seus pacientes
- Entregar uma saúde pública de qualidade.

**Anexo 6 - Persona 1.**

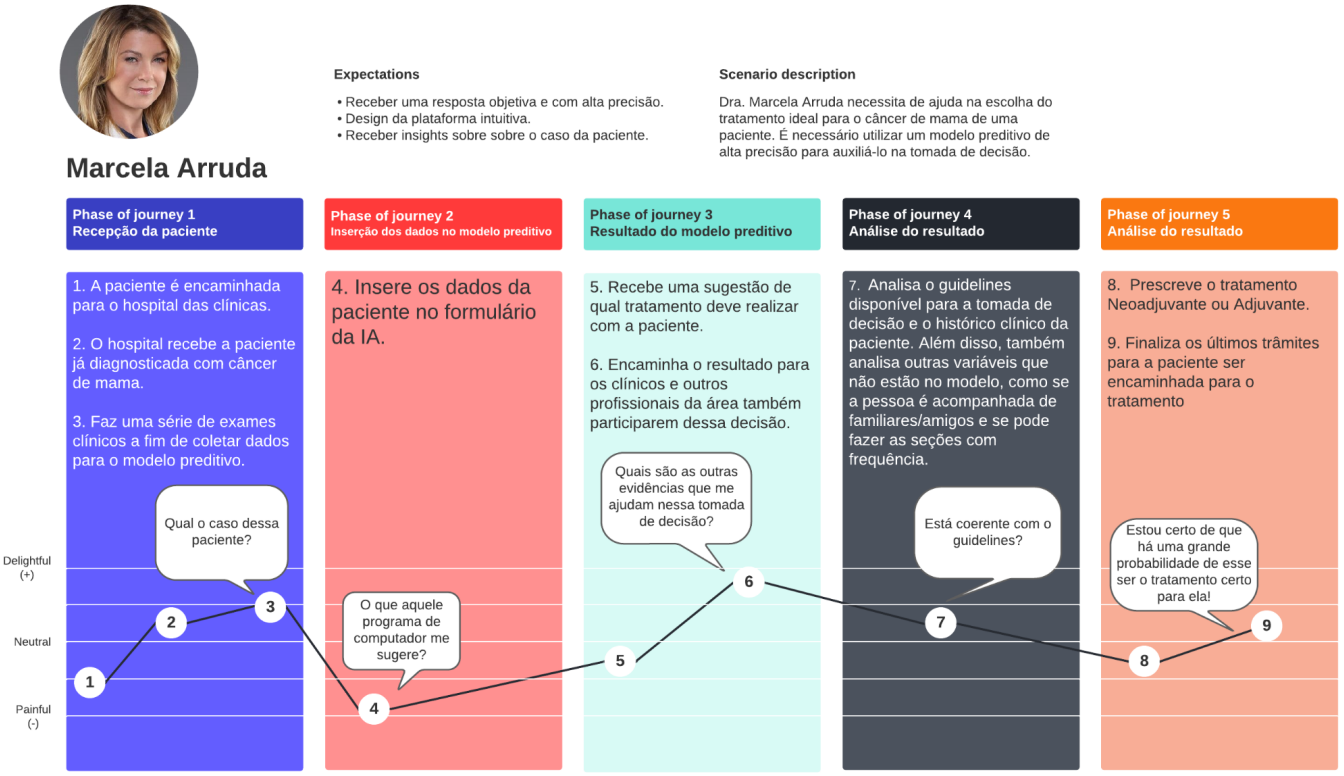
- **Persona 2**

  Paciente impactada pelo modelo preditivo na prescrição do tratamento.


**Anexo 7 - Persona 2.**

## 4.1.7. Jornadas do Usuário

Este mapa fornece a jornada que o usuário terá unicamente com o uso da solução fornecida pela Inteligência Artificial.




**Anexo 8 - Jornada do Usuário.**

### Responsabilidades

1. Marcela Arruda tem a tarefa de obter as informações do paciente e incluí-las no algoritmo de previsão.
2. Além disso, ela precisa executar a simulação, revisar junto com sua equipe a saída do modelo, considerando os fatores que ele não considerou, aconselhar a paciente sobre a opção de tratamento mais adequada e direcioná-la para ele.

## 4.2. Compreensão dos Dados

O modelo preditivo apresentado é baseado em dados provenientes dos prontuários de pacientes do Hospital das Clínicas de São Paulo (HCSP), que foram diagnosticados com câncer de mama em diferentes estágios. Os dados estão disponíveis nos formatos CSV (comma separated values) e XLSX (planilha padrão do Excel).

O banco de dados principal contém 5.000 linhas e 78 colunas, registrando cerca de 4.272 pacientes únicos. Para a solução desenvolvida, foram selecionados conteúdos relevantes dos bancos: Registro de tumor, Peso e altura, Histopatologia e Demográfica.

Devido à dispersão dos dados, é necessário mesclá-los de forma consistente e confiável, utilizando identificadores únicos para cada paciente. Para isso, são necessárias soluções automáticas de limpeza e tratamento de dados. É importante notar que as informações coletadas podem apresentar um viés, já que o HCSP é um prestador de serviços médicos terciários altamente especializados, que recebe pacientes que requerem tratamentos e terapias avançados.

Devido ao tamanho original da base de dados, é inviável utilizá-la em todas as etapas da definição do modelo. Portanto, são selecionados subconjuntos de dados, com foco em parâmetros principais priorizados por ordem de importância para uma análise inicial, os dados das colunas: Tempo desde o primeiro diagnóstico até a recidiva, Combinação dos tratamentos realizados no hospital, Distribuição da combinação de tratamentos, Teve recidiva e Data de recidiva.

### 4.2.1. Exploração dos dados

Iniciando a exploração de dados, importamos todas as bibliotecas necessárias para o desenvolvimento da exploração, tendo como principal utilização o pandas e o numpy. O primeiro passo após a importação das bibliotecas, é começarmos a estruturar a base para sabermos quais são as colunas, tipos, dimensão da tabela e valores únicos.

▾ Etapa 1: Análise Exploratória

▾ Etapa 1.1: Estrutura da base

Entender primeiro a estrutura: quais são as colunas, tipos, dimensão da tabela, valores únicos

```
[ ]  ## Base registro de tumor

     # Checar tipos das colunas
     dfr_raw.dtypes

     # Checar número de linhas e colunas
     dfr_raw.shape

     # Visualizar amostra dos dados
     dfr_raw.sample(10)

     # Contar valores únicos por coluna
     dfr_raw.nunique()

     # Checar percentual de nulos por coluna
     dfr_raw.isnull().mean() * 100
```

```
Record ID                                                  int64
Repeat Instrument                                         object
Repeat Instance                                          float64
Data da primeira consulta institucional [dt_pci]          object
Data do diagnóstico                                       object
```

**Anexo 9 - Colab - Estrutura da base.**

**a) Colunas numéricas e categóricas:**

Para identificação das colunas numéricas e categóricas, acessamos a propriedade "`dtypes`" dos dataframes de Registro de Tumor, Peso e Altura, Histopatologia e Demográfica.

```
# Checar tipos das colunas
dfr_raw.dtypes
```

**Anexo 10 - Colab - Identificação de colunas.**

Em seguida, teremos o retorno de uma lista com todas as colunas do dataframe, indicando seu tipo original, com os tipos “object” para categórico e “float64”/”int64” para numérico. Note que esses são os tipos assumidos na hora da importação. Precisamos checar e fazer as conversões necessárias.

**Para a base Registro de Tumor, temos as seguintes colunas e tipos:**

| | |
|---|---|
| Record ID | Numérico (int) |
| Repeat Instrument | Categórico |
| Repeat Instance | Numérico (float) |
| Data da primeira consulta institucional [dt_pci] | Data |
| Data do diagnóstico | Data |
| Código da Topografia (CID-O) | Categórico |
| Código da Morfologia de acordo com o CID-O | Numérico (float) |
| 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 | Data |
| Combinação dos Tratamentos Realizados no Hospital | Categórico |
| Ano do diagnóstico | Numérico (float) |
| Lateralidade do tumor | Categórico |
| Data de Recidiva | Data |
| Tempo desde o diagnóstico até a primeira recidiva | Numérico (float) |
| 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 |

**Anexo 11 - Colab - Identificação de colunas numéricas e categóricas da base Registro de tumor.**

**Para a base Peso e Altura, temos as seguintes colunas e tipos:**

| | |
|---|---|
| Record ID | Numérico (int) |
| Repeat Instrument | Categórico |
| Repeat Instance | Numérico (float) |
| Data: | Data |
| Peso | Numérico (float) |
| Altura (em centimetros) | Numérico (float) |
| IMC | Numérico (float) |

**Anexo 12 - Colab - Identificação de colunas numéricas e categóricas da base Peso e Altura.**

**Para a base Histopatologia, temos as seguintes colunas e tipos:**

| | |
|---|---|
| Record ID | Numérico (int) |
| Repeat Instrument | Categórico |
| Repeat Instance | Numérico (float) |
| Diagnostico primario (tipo histológico) | Categórico |
| Grau histológico | Numérico (float) |
| Subtipo tumoral | Numérico (float) |
| Receptor de estrogênio | Categórico |
| Receptor de progesterona | Categórico |
| Ki67 (>14%) | Categórico |
| Receptor de progesterona (quantificação %) | Categórico |
| Receptorde Estrogênio (quantificação %) | Categórico |

| | |
|---|---|
| Indice H (Receptorde progesterona) | Numérico (float) |
| HER2 por IHC | Categórico |
| HER2 por FISH | Categórico |
| Ki67 (%) | Numérico (float) |

**Anexo 13 - Colab - Identificação de colunas numéricas e categóricas da baseHispatologia.**

**Para a base Demográfica, temos as seguintes colunas e tipos:**

| | |
|---|---|
| Record ID | Numérica (int) |
| Repeat Instrument | Numérico (float) |
| Repeat Instance | Numérico (float) |
| Escolaridade | Categórico |
| Idade do paciente ao primeiro diagnóstico | Numérico (float) |
| 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 | Categórico |
| Última informação do paciente | Categórico |
| Tempo de seguimento (em dias) - desde o último tumor no caso de tumores múltiplos [dt_pci] | Numérico (float) |
| Já ficou grávida? | Categórico |
| Quantas vezes ficou gráfida? | Numérico (float) |
| Número de partos | Numérico (float) |
| Idade na primeira gestação | Numérico (float) |
| Abortou | Categórico |
| Amamentou na primeira gestação? | Categórico |
| Por quanto tempo amamentou? | Numérico (float) |
| História familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? | Categórico |

| | |
|---|---|
| (choice=Não) | |
| 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 caso) | Categórico |
| Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, apenas 1 caso) | Categórico |
| Historia familiar de câncer relacionado a síndrome de câncer de mama e ovário hereditária? (choice=Sim - 2º grau, mais de 1 caso) | Categórico |
| Idade da primeira mentruação | Numérico (float) |
| 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 | Data |
| 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 | Data |
| Data de início da Radioterapia | Data |

**Anexo 14 - Colab - Identificação de colunas numéricas e categóricas da base Demográfica.**

Como observação, as colunas de tipo data podemos converter para o tipo numérico, como dias a partir de uma data de referência.

**b) Estatística descritiva das colunas.**

Após a análise inicial dos dados e a classificação das colunas numéricas e categóricas, realizamos a estatística descritiva das colunas numéricas:

**Base Registro de Tumor**:

| | repeat instance | Código da morfologia (CID-O) | Tempo desde diagnóstico até primeira recidiva | Teve recidiva |
|---|---|---|---|---|
| contagem | 4677 | 4677 | 1299 | 4677 |
| média | 1,1009 | 84865,3613 | 633,6428 | 0,2777 |
| desv. padrão | 0,3831 | 1115,2255 | 535,3895 | 0,4479 |
| min | 1 | 80103 | 0 | 0 |
| 25% | 1 | 85003 | 254 | 0 |
| 50% | 1 | 85003 | 489 | 0 |
| 75% | 1 | 85003 | 867,5 | 1 |
| max | 8 | 99873 | 3462 | 1 |

**Anexo 15 - Colab - Estatística descritiva das colunas numéricas da base Registro de tumor.**

**Base Peso e Altura:**

| | repeat instance | Peso | Altura (cm) | IMC |
|---|---|---|---|---|
| contagem | 51382 | 45178 | 49928 | 49919 |
| média | 11,4986 | 71,2374 | 157,1957 | 24,8454 |
| desv. padrão | 10,1454 | 241,738 | 7,234 | 10,7376 |
| min | 1 | 1 | 0 | 0 |
| 25% | 4 | 59,65 | 152 | 22,8 |
| 50% | 9 | 68,35 | 157 | 26,9 |
| 75% | 16 | 78,6 | 162 | 30,8 |
| max | 96 | 51350 | 191 | 347,7 |

**Anexo 16 - Colab - Estatística descritiva das colunas numéricas da base Peso e Altura.**

**Base de Histopatologia:**

| | repeat instance | Grau histológico | Subtipo tumoral | Índice h receptor de progesterona | KI67 % |
|---|---|---|---|---|---|
| contagem | 4794 | 1467 | 4695 | 592 | 3867 |
| média | 1,1358 | 2,257 | 2,958 | 197,0895 | 36,5343 |
| desv. padrão | 0,636 | 0,6535 | 1,2967 | 98,4853 | 24,556 |
| min | 1 | 1 | 1 | 0 | 0 |
| 25% | 1 | 2 | 2 | 120 | 18 |
| 50% | 1 | 2 | 3 | 240 | 30 |
| 75% | 1 | 3 | 4 | 285 | 50 |
| max | 17 | 3 | 5 | 300 | 100 |

**Anexo 17 - Colab - Estatística descritiva das colunas numéricas da base Hispatologia.**

**Base Demográfica:**

| | Idade paciente primeiro diagnóstico | Tempo seguimento desde último tumor (tumores múltiplos DT PCI) | Quantas vezes ficou grávida |
|---|---|---|---|
| contagem | 4092 | 4270 | 44 |
| média | 54,2478 | 1475,0037 | 2,3182 |
| desv. padrão | 13,5741 | 859,6224 | 1,4105 |
| min | 22 | 0 | 1 |
| 25% | 45 | 956,25 | 1 |
| 50% | 54 | 1282 | 2 |
| 75% | 64 | 1817,75 | 3 |
| max | 98 | 4503 | 7 |

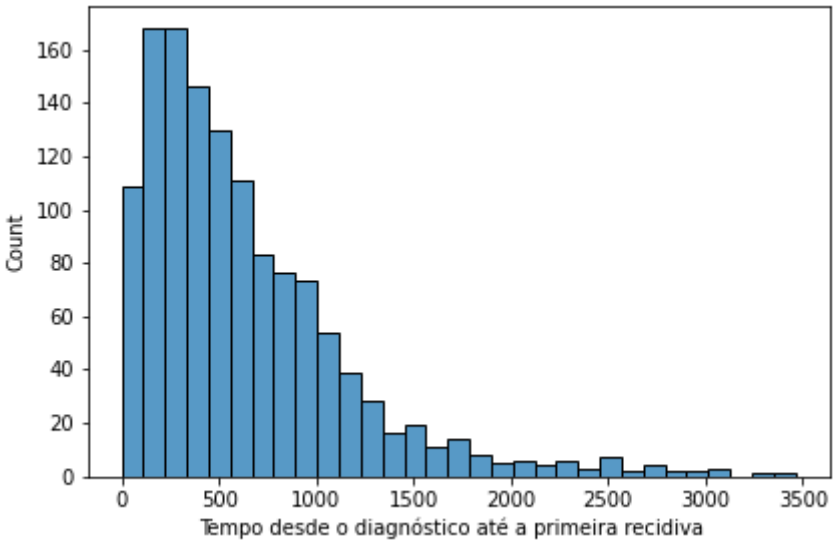
| | Número partos | Idade primeira gestação | Por quanto tempo amamentou | Idade primeira menstruação |
|---|---|---|---|---|
| contagem | 2 | 897 | 688 | 1025 |
| média | 1,5 | 23,058 | 19,0436 | 12,8917 |
| desv. padrão | 0,7071 | 5,6652 | 23,1051 | 2,1044 |
| min | 1 | 0 | 0 | 0 |
| 25% | 1,25 | 19 | 6 | 12 |

| | | | | |
|---|---|---|---|---|
| 50% | 1,5 | 22 | 12 | 13 |
| 75% | 1,75 | 26 | 24 | 14 |
| max | 2 | 53 | 260 | 37 |

**Anexo 18 - Colab - Estatística descritiva das colunas numéricas da base Demográfica.**

### c) Visualização de Gráficos

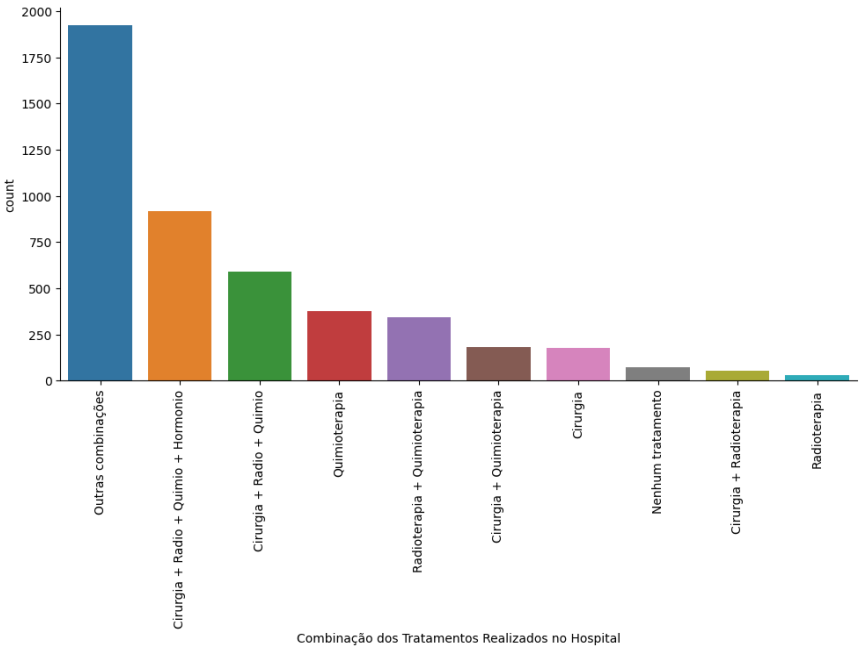
**Gráfico 1:**




**Anexo 19 - Colab - Tempo desde o diagnóstico até a primeira recidiva**

O gráfico demonstra que o tempo decorrido até a primeira recidiva desde o diagnóstico é predominantemente concentrado nos primeiros 500 dias.

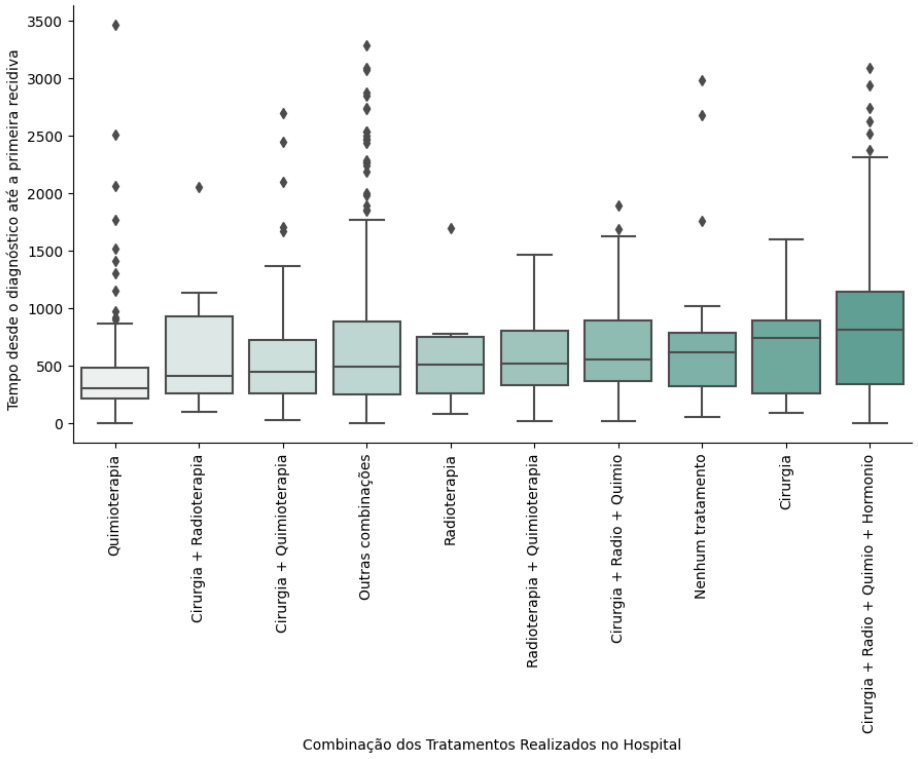
**Gráfico 2:**




**Anexo 20 - Colab - Combinação dos tratamentos realizados no hospital**

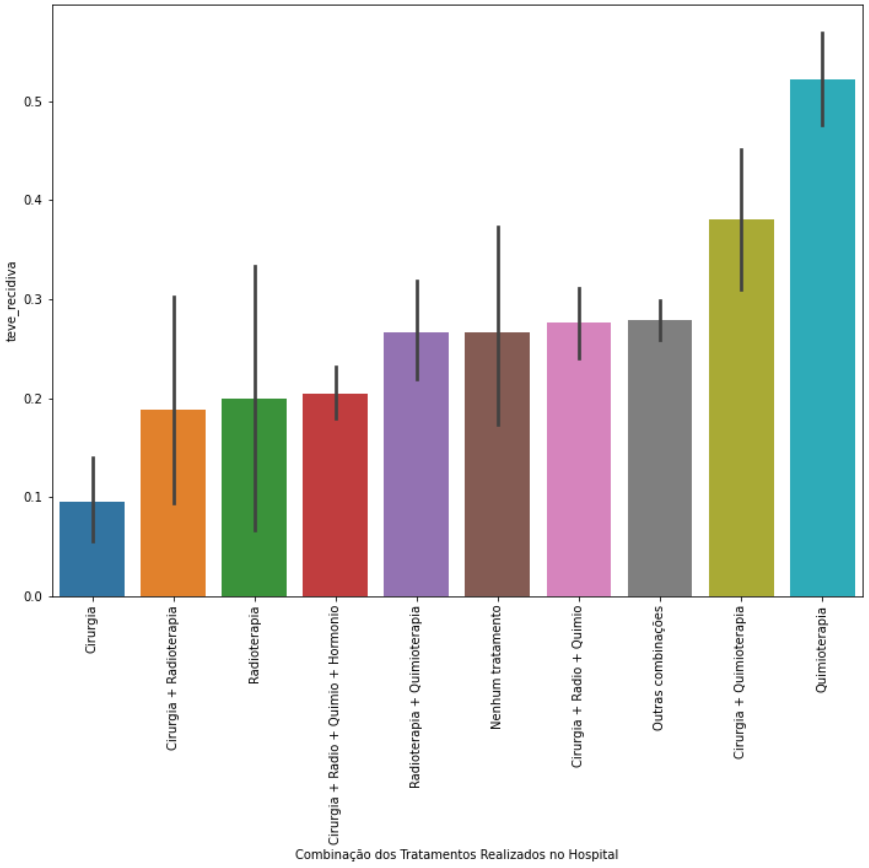
Podemos observar que o hospital das clínicas possui um maior índice de tratamento utilizando outras combinações.

**Gráfico 3:**




**Anexo 21 - Colab - Combinação dos tratamentos realizados no hospital vs Tempo desde diagnóstico até a primeira recidiva.**

**Gráfico 4:**




**Anexo 22 - Colab - Combinação dos tratamentos realizados no hospital.**

Ao observar o gráfico, é possível notar que os pacientes que iniciaram o tratamento com quimioterapia apresentaram o maior número de recidivas, enquanto aqueles que iniciaram o tratamento com cirurgia não foram tão afetados.

## 4.2.2. Pré-processamento dos dados:

A  limpeza dos dados é importante antes do tratamento para impedir que valores nulos, inválidos, duplicados, ausentes ou inconsistentes possam afetar a acuracidade das análises e evitar problemas advindos de processos imprecisos como a existência dos missing values ou outliers.

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

Os outliers foram identificados durante o tratamento dos dados, a primeira tabela analisada foi a de dados demográficos  e identificamos outliers nas colunas “idade_primeiro_diagnostico”, “tempo_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_tumores_multiplos_dt_pci”, “quantas_vezes_ficou_gravida”, “idade_primeira_gestacao”, “tempo_de_amamentacao ”, “idade_primeira_menstruacao”.

Antes de tratarmos cada coluna individualmente, fizemos a identificação dos outliers nas colunas numéricas.

- Identificação de outliers para as colunas numéricas:

```
#estamos usando a média +- 3 desvio padrão para identificar outliers , supondo que a distribuicao é normal
numericas = ['idade_primeiro_diagnostico', 'tempo_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_tumores_multiplos_dt_pci','quantas_vezes_ficou_grav:
for colu in numericas:
  mean = dfd[colu].mean()
  std = dfd[colu].std()
  higher_limit = mean + (3 * std)
  lower_limit = mean - (3 * std)
  outliers = dfd[['record_id',colu]][dfd[colu]>higher_limit ]
  outliers2 = dfd[['record_id',colu]][dfd[colu]<lower_limit ]
  print(f'na coluna {colu} o outlier acima é: \n {outliers} \n o outlier abaixo é: \n {outliers2} \n')
```

```
na coluna idade_primeiro_diagnostico o outlier acima é:
      record_id  idade_primeiro_diagnostico
2917      61384                        97.0
3317      68212                        95.0
3407      69335                        97.0
3932      74814                        98.0
4046      78581                        96.0
```

**Anexo 23 - Colab -  média x desvio padrão.**

Após a identificação das colunas com outliers, fizemos a análise:

**Coluna 1**: “idade_primeiro_diagnostico”, substituímos todos os valores acima de 94 pela moda como mostra o código abaixo.

```
dfd['idade_primeiro_diagnostico'] = np.where((dfd['idade_primeiro_diagnostico'] > 94), idade_diag_moda , dfd['idade_primeiro_diagnostico'])
```

**Anexo 24 - Colab - primeiro diagnóstico.**

**Coluna 2**:
“tempo_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_tumores_multiplos_dt_pci”, substituimos todos os valores em dias que somavam mais de cinco anos entre a recidiva para o valor em dias equivalente a três anos.

Tratamento dos outliers da coluna "tempo_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_tumores_multiplos_dt_pci":

```
[ ]  _tumor_no_caso_tumores_multiplos_dt_pci'] > 2347 ), 1407 , dfd['tempo_seguimento_em_dias_desde_o_ultimo_tumor_no_caso_tumores_multiplos_dt_pci'])
```

**Anexo 25 - Colab - tratamento de outlier 1.**

**Coluna 3:** “idade_da_primeira_gestacao, todas as pessoas registradas com a idade da primeira gestação igual a zero, utilizamos a moda dos registros para definir o valor mais próximo.

Tratamento dos outliers da coluna "idade_primeira_gestacao":

```
[ ]  'idade_primeira_gestacao'] > 40 ) | dfd['idade_primeira_gestacao'] == 0 , dfd['idade_primeira_gestacao'].mode() , dfd['idade_primeira_gestacao'])
```

**Anexo 26 - Colab - tratamento de outlier 2.**

**Coluna 4:** tempo_de_amamentacao, em todos os casos que o tempo de amamentação registrado tenha sido maior que vinte e cinco meses, o valor será substituído pela moda.

Tratamento dos outliers da coluna "tempo_de_amamentacao":

```
[ ]  dfd['tempo_de_amamentacao'] = np.where((dfd['tempo_de_amamentacao'] > 25 ) , dfd['tempo_de_amamentacao'].mode() , dfd['tempo_de_amamentacao'])
```

**Anexo 27 - Colab - tratamento de outlier 3.**

**Coluna 5:** “idade_primeira_mentruacao”, em todos os casos em que a idade da primeira menstruação esteja registrada acima de 19 anos o valor será substituído para a moda.


**Anexo 28 - Colab - tratamento de outlier 4.**

### 4.2.3. Hipóteses:

a) **Ki67 está relacionado com subtipo tumoral e grau histológico, influenciando no tipo e no grau do tumor.**

A hipótese de que o ki67 está relacionado com o subtipo tumoral e o grau histológico sugere que a expressão de ki67 pode ser um indicador da agressividade e potencial de crescimento do tumor. O ki67 é uma proteína que é expressa durante a fase ativa do ciclo celular e é frequentemente usada como um marcador para medir a taxa de proliferação celular em tecidos tumorais. A expressão de ki67 pode ser um indicador de como rapidamente as células do tumor estão se dividindo e se replicando.

A correlação entre expressão de ki67 e o subtipo tumoral pode ajudar a diferenciar os tumores em categorias distintas, o que pode ter implicações para o prognóstico e tratamento do paciente. Alguns subtipos tumorais podem ser mais agressivos e resistentes ao tratamento do que outros, então a capacidade de identificar esses subtipos pode ajudar os médicos a selecionar os tratamentos mais adequados para seus pacientes.

Além disso, a correlação entre a expressão de ki67 e o grau histológico do tumor sugere que o ki67 pode ser um indicador do nível de diferenciação celular presente no tecido tumoral. Quanto menos diferenciadas as células do tumor são, maior a probabilidade de que o tumor seja mais agressivo e mais difícil de tratar. Portanto, a análise da expressão de ki67 pode ser útil na determinação do prognóstico e do tipo de tratamento adequado para o paciente.

b) **Pacientes que possuem tumores múltiplos tem uma recidiva maior.**

A presença de múltiplos tumores pode ser indicativa de uma doença mais avançada ou agressiva, o que aumenta a probabilidade de recidiva após o tratamento inicial. Além disso, a presença de tumores múltiplos pode indicar que o paciente tem uma maior predisposição genética ou ambiental para desenvolvimento de tumores, o que também pode aumentar o risco de recidiva.

Outro fator a considerar é que a presença de múltiplos tumores pode tornar mais difícil a remoção completa do tecido tumoral durante a cirurgia. Isso pode deixar células tumorais remanescentes no corpo, o que aumenta o risco de recidiva.

Além disso, pacientes com tumores múltiplos geralmente exigem tratamentos mais agressivos, como a cirurgia em mais de uma área ou radioterapia em várias áreas, o que

pode aumentar o risco de complicações e efeitos colaterais que podem afetar negativamente a resposta do paciente ao tratamento.

Em resumo, a presença de tumores múltiplos pode indicar uma doença mais avançada ou agressiva, maior predisposição para desenvolver tumores, dificuldades na remoção completa do tecido tumoral durante a cirurgia e a necessidade de tratamentos mais agressivos, todos os quais podem aumentar o risco de recidiva do câncer.

**c) Recidivas são mais comuns em pacientes que não passaram por cirurgia.**

A cirurgia é muitas vezes uma forma eficaz de remover completamente o tecido tumoral. Quando todo o tecido tumoral é removido, a probabilidade de que células cancerosas remanescentes cresçam novamente é menor, o que pode reduzir o risco de recidiva. Pacientes que não passaram por cirurgia podem ter um tecido tumoral residual que pode aumentar o risco de recidiva.

Além disso, a cirurgia pode ser combinada com outras modalidades de tratamento, como quimioterapia ou radioterapia, para maximizar a eficácia do tratamento. A combinação dessas modalidade de tratamento pode ajudar a destruir quaisquer células cancerosas remanescentes que possam estar presentes no corpo, reduzindo assim o risco de recidiva.

Por outro lado, pacientes que não passaram por cirurgia podem depender exclusivamente de outras modalidades de tratamento, como quimioterapia ou radioterapia, que nem sempre conseguem remover todo o tecido tumoral ou matar todas as células cancerosas. Isso pode aumentar a probabilidade de células cancerosas remanescentes crescerem novamente, resultando em uma maior taxa de recidiva.

## 4.2.4. Política de privacidade LGPD:

**POLÍTICA DE PRIVACIDADE ACERCA DO PROJETO AGATHA**

A equipe Hígia é responsável pelo desenvolvimento deste modelo preditivo.

O projeto Agatha é comprometido em proteger a privacidade e segurança dos dados pessoais dos usuários. Este documento descreve como coletamos, armazenamos e utilizamos informações pessoais relacionadas à saúde dos usuários.

**Coleta de dados:**

Nós coletamos os dados necessários para o desenvolvimento do modelo preditivo de forma legal, justa e transparente, os dados coletados podem ser relacionados à saúde dos usuário, quando é necessário para fornecer os serviços oferecidos pelo projeto. Estas informações incluem, mas não estão limitadas a, informações médicas, histórico de saúde e informações de contato como, e-mail e número de telefone .

**Armazenamento de dados:**

Os dados pessoais coletados são armazenados em servidores seguros e só são acessíveis por pessoas autorizadas que precisam dessas informações para desempenhar suas funções.

**Uso de dados:**

Os dados pessoais dos usuários são utilizados apenas para fornecer os serviços oferecidos pelo projeto e para fins de pesquisa e análise interna. Não compartilhamos estas informações com terceiros, exceto se exigido por lei ou se for necessário para prestar os serviços oferecidos pelo projeto.

**Segurança de dados:**

Tomamos medidas de segurança razoáveis para proteger os dados pessoais dos usuários contra perda, mau uso, acesso não autorizado, alteração e destruição.

**Alterações na política de privacidade:**

Reservamo-nos o direito de alterar esta política de privacidade a qualquer momento. Qualquer alteração será publicada em nosso site.

**Cookies**

Nós usamos cookies para melhorar a experiência do usuário e personalizar o conteúdo exibido. Os cookies também são usados para coletar informações anônimas sobre a navegação do usuário em nosso site. Os usuários podem controlar o uso de cookies nas configurações do navegador.

**Contato**:

Se tiver alguma dúvida sobre esta política de privacidade ou sobre como tratamos os seus dados pessoais, entre em contato conosco no email: hígia@sou.inteli.edu.br

## 4.3. Preparação dos Dados e Modelagem

**1. Modelo supervisionado:**

**a) Modelagem para o problema (proposta de features com a explicação completa da linha de raciocínio).**

Fizemos a seleção das *features* com base em sua correlação com o nosso objetivo, que foi inicialmente definido como sucesso quando o paciente não sofreu recidiva ou não faleceu devido ao câncer, e fracasso quando houve recidiva ou morte por câncer. No entanto, reconhecemos que nosso objetivo ainda apresenta limitações, pois não considera uma série de outros fatores, como tempo de sobrevida e local de recidiva. Planejamos aprimorá-lo ao longo das próximas *sprints*.

É fundamental destacar que nosso modelo não prevê o tipo de tratamento, mas sim o sucesso ou fracasso dele. Uma das características relevantes é o próprio tratamento, o qual será testado duas vezes: primeiro com o adjuvante e depois com o neoadjuvante. Em seguida, analisaremos qual deles obteve maior sucesso. Caso ambos tenham resultados em fracasso, iremos calcular a distância euclidiana dos nós previstos em relação ao plano que divide as possíveis classificações, e a indicação será o mais próximo desse plano.

Em outras palavras, estamos selecionando as *features* com base na sinergia entre o target e uma determinada coluna. No entanto, podemos quebrar essa regra se identificarmos por meios teóricos que há uma relação entre elas. Nesse caso, essa coluna será adicionada como *feature*.

**b) Métricas relacionadas ao modelo:**

- Para obter a melhor precisão possível, iremos utilizar as seguintes métricas:
    - **Acurácia:** essa medida é utilizada para avaliar a porcentagem de previsões corretas nos conjuntos de teste e treino. No nosso caso, a acurácia irá mensurar a quantidade de pacientes que foram classificados corretamente, ou seja, aqueles que não tiveram recidiva nem faleceram devido ao câncer.
    - **Recall:** essa medida avalia a proporção de casos positivos que foram corretamente identificados pelo modelo. No contexto do projeto 'Agatha', ela representa a quantidade de pacientes previstos como tendo sucesso, ou seja, aqueles que não faleceram nem apresentaram recidiva, que foram identificados corretamente.
    - **Especificidade:** essa medida refere-se à quantidade de casos negativos reais que foram identificados pelo modelo. No contexto do projeto 'Agatha', esse valor seria a quantidade de pacientes previstos como tendo fracasso, ou seja, aqueles que morreram ou apresentaram recidiva, que realmente foram afetados por esses eventos.

- **AUC-ROC:** essa medida fornece uma avaliação geral do modelo em intervalos específicos de classificação. No contexto do projeto, ela se refere à capacidade do modelo de prever corretamente os casos em que ocorreram recidiva ou morte, bem como os casos em que nenhum desses eventos ocorreu.

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

O modelo SVM, ou Support Vector Machine em inglês, é o primeiro modelo candidato a ser utilizado no projeto. Ele tem como objetivo encontrar um hiperplano de inúmeras dimensões que permita uma classificação distinta e precisa dos dados. Esse modelo é amplamente utilizado em problemas de classificação, pois busca maximizar a margem entre as diferentes classes, o que contribui para uma melhor capacidade de generalização e previsão de novos dados.

O modelo busca encontrar o hiperplano que melhor separa as diferentes classes de dados, através da maximização da margem, ou seja, da maior distância entre os pontos mais próximos de cada classe. Isso permite uma maior precisão na classificação de novos dados, já que a margem maior implica em um menor risco de sobreajuste (overfitting) aos dados de treinamento e, consequentemente, uma melhor capacidade de generalização do modelo.

Esse modelo foi escolhido porque estamos lidando com dois tipos possíveis de classificação (sucesso ou fracasso no tratamento) e porque priorizamos a precisão na recomendação do melhor tipo de tratamento para o câncer de mama aos pacientes.

Com base nas métricas selecionadas, o modelo apresentou um desempenho satisfatório até o momento, com valores de 70% para acurácia de treino, 72% para acurácia de teste, 58% de sensibilidade, 99% de especificidade e 53% na AUC-ROC. Esses resultados indicam que o modelo possui um bom potencial e pode ser aprimorado nas próximas versões.

## 4.4. Comparação de Modelos

**a) Escolha da métrica e justificativa.**

Entendemos que, para realizar comparações entre modelos, é essencial utilizar métricas comuns no mundo da medicina. Por esse motivo, escolhemos utilizar as seguintes métricas de classificação para avaliar a eficácia do modelo:

- **Acurácia:** é uma medida usada para avaliar a porcentagem de previsões corretas em conjuntos de teste e treinamento. No contexto de avaliação de pacientes com câncer de mama, a precisão mede a proporção de pacientes que foram classificados corretamente, ou seja, aqueles que não sofreram recidiva ou faleceram devido ao câncer. É uma medida importante para avaliar a eficácia de um modelo preditivo e sua capacidade de identificar pacientes em risco de desenvolver complicações.
- **Recall:** é uma medida que avalia a proporção de casos positivos que foram corretamente identificados pelo modelo. No contexto do projeto, ela representa a quantidade de pacientes previstos como tendo sucesso, ou seja, aqueles que não faleceram nem apresentaram recidiva e que foram identificados corretamente pelo modelo. A sensibilidade é uma métrica importante para avaliar a capacidade do modelo em identificar pacientes que estão realmente em risco de complicações e, portanto, é uma medida fundamental para avaliar a eficácia do modelo em casos de saúde.
- **Especificidade:** é uma medida que refere-se à proporção de casos negativos reais que foram identificados corretamente pelo modelo. No projeto, essa métrica representa a quantidade de pacientes previstos como tendo fracasso, ou seja, aqueles que morreram ou apresentaram recidiva e que foram identificados corretamente pelo modelo. A especificidade é uma métrica importante para avaliar a capacidade do modelo em identificar pacientes que não estão em risco de complicações e, portanto, é uma medida fundamental para avaliar a eficácia do modelo em casos de saúde.
- **AUC-ROC:** é uma métrica que fornece uma avaliação geral do modelo em intervalos específicos de classificação. No projeto, ela se refere à capacidade do modelo de prever corretamente os casos em que ocorreram recidiva ou morte, bem como os casos em que nenhum desses eventos ocorreu. É importante para avaliar a capacidade geral do modelo de distinguir entre casos positivos e negativos, e é amplamente utilizada em aplicações de machine learning para classificação de dados.

**b) Modelos otimizados:**

**SVM:**
O modelo SVM foi otimizado com o Grid Search, levando em consideração três hiperparâmetros. O primeiro é o C, que pondera o modelo em evitar a classificação errônea dos exemplos de treino através de um peso. Quanto maior o valor de C, maior será a 'punição' para o modelo. O segundo hiperparâmetro é o Kernel, responsável pelo formato da função que o modelo utiliza para dividir melhor os dados. O Kernel pode ser baseado em uma função sigmoidal, polinomial ou radical. Por fim, o último hiperparâmetro é o gamma, que é o coeficiente da função kernel. Quanto maior o valor de gamma, mais precisa a função será.

**KNN:**

O hiperparâmetro da modelagem KNN foi ajustado com o Grid Search, importando a biblioteca 'Grid Search CV' da SKlearn. O parâmetro considerado para o modelo KNN foi o 'n_neighbors', que determina o número de vizinhos mais próximos que o algoritmo utiliza para classificar novos dados. Criamos um intervalo (k_range) inicialmente definido de 1 a 100, mas posteriormente testamos de 1 a 500. No entanto, o resultado foi o mesmo. Para otimizar o código, mantivemos 'k_range = range (1.100)', que permite ao Grid Search testar cada valor nesse intervalo.

Em seguida, utilizamos 'grid.best_params_' e 'grid.best_score_' para retornar, respectivamente, o melhor número de vizinhos com base em algumas métricas, incluindo a acurácia, e a acurácia para a quantidade de vizinhos selecionados. O Grid Search retornou que 'n_neighbors = 64' e o 'best_score' foi de aproximadamente 74%.

**Random Forest:**

Neste modelo, utilizamos o RandomizedSearchCV, uma técnica de busca aleatória, para otimizar os hiperparâmetros do modelo. Foram definidos diferentes valores para os hiperparâmetros "max_depth", "max_features", "min_sample_leaf", "min_samples_split" e "n_estimators". O RandomizedSearchCV buscou aleatoriamente em todas as combinações possíveis de hiperparâmetros dentro do espaço definido e selecionou a melhor combinação de acordo com a métrica de acurácia.

Algumas explicações: o parâmetro "n_inter" definiu o número de iterações que o algoritmo deveria fazer, "cv" determinou o número de folds para validação cruzada, e "n_jobs" estipulou o número de tarefas executadas em paralelo.

A partir dos resultados da hiper parametrização, concluímos que o modelo apresentou um bom desempenho na classificação da classe 1, mas precisa de melhorias na classificação na classe 0. Isso ocorre porque a classe 1 é mais representativa do que a classe 0 no conjunto de dados, o que indica um desbalanceamento do nosso target.

**Logistic Regression:**

Logistic Regression é utilizado para prever a probabilidade de um evento binário ocorrer, ou seja, que pode ter apenas duas opções, com base em um conjunto de features.

Em nosso modelo, utilizamos Logistic Regression com 'param_grid' que é comumente usado para ajustar hiperparâmetros em modelos de machine learning, como modelos de classificação linear com regularização L1 ou L2.

O dicionário 'param_grid' tem duas chaves: 'c' é um parâmetro de regularização que controla a força da penalização para os coeficientes do modelo. Valores menores que 'c' geram modelos mais regularizados e valores maiores de 'c' geram modelos menos regularizados. No modelo, o valor de 'c' é definido como uma lista de 5 valores diferentes: 0.001, 0.01, 0.1, 1 e 10.

Já a segunda chave 'penalty' é um parâmetro que especifica o tipo de regularização aplicado ao modelo. Pode ser L1, que é uma regularização que incentiva a esparsidade dos coeficientes, ou L2, que penaliza os coeficientes de forma mais suave. No modelo, o valor 'penalty' é definido como uma lista com os valores 'l1' e 'l2'.

A classe 'GridSearchCV', irá avaliar todas as combinações possíveis de valores de hiperparâmetros, treinar o modelo para cada combinação e retornar o melhor modelo encontrado.

O parâmetro 'refit=True' indica que o modelo deve ser reajustado com o conjunto de dados completo após a busca em grade ter sido concluída, usando os melhores valores de hiperparâmetros encontrados.

Juntos, esses parâmetros permitem que sejam testadas diferentes combinações de hiperparâmetros no processo de ajuste do modelo, de modo que se possa escolher aqueles que geram o melhor desempenho de acordo com alguma métrica de avaliação como acurácia, sensibilidade, especificidade e etc.

**Naive Bayes:**

Os hiperparâmetros são ajustes que podem ser feitos nos algoritmos para melhorar seu desempenho em diferentes conjuntos de dados. Por meio do processo de ajuste de hiperparâmetros, é possível testar diferentes combinações de valores e avaliar seu desempenho com base em métricas como acurácia, sensibilidade e especificidade. É possível selecionar os valores de hiperparâmetros que geram o melhor desempenho para um determinado conjunto de dados. Isso torna os hiperparâmetros uma parte importante do processo de modelagem de Naive Bayes.

O Naive Bayes é um algoritmo de aprendizado de máquina que se refere às escolhas feitas ao criar o modelo, existem três tipos principais: Bernoulli, Multinomial e Gaussiano.Eles se diferem

pela maneira de tratar o conjunto de entrada, sendo o Bernoulli mais utilizado para tratar dados binários, o Multinomial para tratar dados discretos e o Gaussiano para tratar dados contínuos.

Para entendermos melhor qual tem uma aplicabilidade mais eficiente para o projeto, desenvolvemos o hiperparâmetros dos três tipos. Utilizamos o Grid Search para a modelagem e consideramos os parâmetros de cada, para o Bernoulli o parâmetro 'alpha' com os valores [0, 1.0e-10, 1.0, 1.3] e o parâmetro 'fit_prior' com os valores[True, False], para o Multinomial o parâmetro 'alpha' com os valores [0, 0.2, 1.0, 1.3], e os parâmetros 'force_alpha' e 'fit_prior' com os valores, [True, False], e para a Gaussiano o parâmetro 'var_smoothing', com os valores [1e-12,1e-11,1e-10,1e-9, 1e-8, 1e-7, 1e-6,1e-5,1e-4,1e-3,1e-2]

**c) Definição do modelo escolhido e justificativa.**

No projeto Agatha, optamos pelos modelos SVM e Logistic Regression como finalistas na definição, pois ambos apresentaram resultados promissores e mais robustos do que os outros modelos avaliados. No entanto, estamos enfrentando um problema de desbalanceamento de dados, o que nos impede de escolher definitivamente o modelo final do projeto. Nessa decisão levaremos em conta tanto a precisão do modelo quanto a curva ROC-AUC, garantindo que o modelo escolhido seja o mais robusto possível para a recomendação do tratamento do câncer de mama.

## 4.5. Avaliação

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

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

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

# 5. Conclusões e Recomendações

*Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção "Anexos".*

*Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.*

# 6. Referências

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

*Um exemplo de referência de livro:*

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

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

# Anexos

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