Modelo preditivo da variabilidade da evolução do câncer de mama - Faculdade de Medicina da Universidade de São Paulo

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
04/08/2022	Gabriela Barretto, Pedro Romão e Wagner Estevam	1.0	Criação do documento.
10/08/2022	Gabriela Barretto e Wagner Estevam	1.5	Preenchimento das seções 1, 2, 4.1 e 4.2.
14/08/2022	Elias Biondo	1.9	Refinamento geral pré-entrega.

Sumário

1. Introdução **4**

2. Objetivos e Justificativa **5**

2.1. Objetivos 5

2.2. Proposta de solução 5

2.3. Justificativa **5**

3. Metodologia **6**

3.1. CRISP-DM 6

3.2. Ferramentas 6

3.3. Principais técnicas empregadas 6

4. Desenvolvimento e Resultados **7**

4.1. Compreensão do Problema 7

4.1.1. Contexto da indústria 7

4.1.2. Análise SWOT **9**

4.1.3. Planejamento Geral da Solução 9

4.1.4. Value Proposition Canvas **10**

4.1.5. Matriz de Riscos 10

4.1.6. Personas 11

4.1.7. Jornadas do Usuário 11

4.2. Compreensão dos Dados 12

4.3. Preparação dos Dados 14

4.4. Modelagem 15

4.5. Avaliação 16

4.6 Comparação de Modelos 17

5. Conclusões e Recomendações **18**

6. Referências **19**

Anexos **20**

1. Introdução

A Universidade de São Paulo (USP) é a melhor universidade da América Latina de acordo com o ranking Best Global Universities (2021), recebendo destaque na área de pesquisa científica nacional e internacional. É, também, a melhor universidade pública do país. A Faculdade de Medicina da Universidade de São Paulo (FM-USP) corresponde ao principal parceiro de negócio desse projeto. Associada ao Hospital das Clínicas (HC/FM-USP), uniram forças para a criação de uma solução inovadora. Posicionando-se sempre à favor da vida, sua área de atuação, no contexto desse produto, é a saúde humana e a oncologia (especialidade médica que estuda os cânceres e a forma como essas doenças se desenvolvem).

A evolução do câncer de mama e suas respostas a tratamentos convencionais é muito variável, o que faz com que médicos não saibam tratar, de forma objetiva, casos particulares e, ainda, identificar o grau de risco dos pacientes e o seu tempo de sobrevida estimado. Torna-se interessante, a partir disso, o surgimento de alguma forma de tecnologia para estimar essas variáveis visando uma distribuição mais eficaz e eficiente dos recursos públicos do Sistema Único de Saúde (SUS). Dessa maneira, além da economia de aportes, vidas poderão ser salvas através de um tratamento direcionado e personalizado.

2. Objetivos e Justificativa

2.1. Objetivos

O principal objetivo do parceiro de negócio é viabilizar a criação de uma solução capaz de analisar, filtrar e classificar dados de pacientes com câncer a fim de identificar a variabilidade da evolução da doença, bem como as possíveis respostas a tratamentos já implantados, para, assim, detectar padrões que indiquem aos profissionais de saúde uma possível estimativa de risco e grau de sobrevida de um paciente.

Como objetivo secundário, destaca-se o direcionamento de recursos disponíveis de maneira mais eficaz e eficiente, explica-se: com a identificação de padrões que indiquem aos profissionais de saúde uma possível estimativa de risco e grau de sobrevida de um paciente, será possível conduzir menos ou mais consultas e exames a depender da necessidade do paciente, tornando o processo totalmente personalizável às variáveis identificadas.

Não fica de fora, também, o aumento da qualidade de vida dos enfermos, uma vez que eles poderão se deslocar menos ou planejar mais idas ao hospital dependendo de suas condições clínicas.

2.2. Proposta de solução

Dado o exposto, a partir da matéria-prima disponibilizada pelo parceiro - os dados de pacientes com câncer - a proposta de solução da equipe é um modelo preditivo da variabilidade da evolução do câncer de mama, isto é, um modelo de aprendizado de máquina e inteligência artificial que identifique padrões da variabilidade da evolução da doença em relação aos tratamentos e terapias aplicados sob os enfermos, a fim de retornar um prognóstico com padrões explícitos que indiquem aos profissionais de saúde o risco e grau de sobrevida de um paciente específico com base em seus dados clínicos.

2.3. Justificativa

Um algoritmo de análise preditiva, como proposto, assegurará a consideração de todos os dados relevantes à variabilidade da evolução da doença. Apesar de ainda pouco explorados, esses algoritmos possuem a capacidade de mudar como a humanidade pensa e executa a medicina nos tempos atuais. Essas análises nunca poderiam ser feitas, de maneira generalizada, por seres humanos, uma vez que a quantidade de dados existentes é extremamente massiva. Tendo como premissa básica a qualidade da coleta dos dados disponibilizados e a ausência de viéses no banco, é garantida uma boa assertividade nas predições do algoritmo, viabilizando, dessa maneira, um sistema completo e rápido de análise de casos de pacientes com a doença e o seu direcionamento eficiente.

3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Colaboratory)

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

Não foram identificados concorrentes diretos do parceiro de negócio, uma vez que o mesmo se posiciona como uma instituição pública a favor da sociedade e em serviço por ela, explica-se: toda produção de conhecimento interna é passível de uso por qualquer outra organização, desde que sejam respeitados os devidos direitos de uso e direitos comerciais. No entanto, é possível destacar algumas outras empresas que atuam na mesma fatia de mercado, como por exemplo:

- O Hospital A.C.Camargo, especializado no diagnóstico, tratamento e pesquisa de câncer em humanos;
- O Instituto de Câncer do Estado de São Paulo (ICESP), maior centro oncológico da América Latina, especializado no tratamento de casos de câncer de alta complexidade;
- O Instituo Nacional de Câncer (INCA), órgão auxiliar do Ministério da Saúde no desenvolvimento e coordenação das ações integradas para a prevenção e o controle de câncer no Brasil

Como uma instituição pública sem fins lucrativos, fica evidente o modelo de negócio do parceiro quanto ao seu pioneirismo e excelência nos âmbitos de ensino, pesquisa e extensão universitária, além da modernização e inovação tecnológica. Não cedendo a pressões do mercado, a instituição é capaz de, por meio de investimentos da União, inovar e estar na vanguarda da ciência, sempre com as últimas e mais modernas soluções de problemas.

Sobre o uso crescente de algoritmos de aprendizado de máquina e inteligência artificial, emerge-se a tendência da substituição ou o aumento das capacidades humanas para a compreensão objetiva de conceitos científicos, matemáticos e humanos. Em 2021, o setor de saúde digital alcançou um recorde de US$ 57,2 bilhões de dólares segundo os dados do relatório State of Digital Health da plataforma norte-americana de inteligência de mercado CB insights, correspondendo um aumento de 79% em relação ao ano anterior.

Em uma visão macro, segue abaixo a análise da indústria realizada de acordo com as diretrizes de Michael Porter e suas Cinco Forças:

1. **Rivalidade entre os concorrentes:** por ser uma instituição governamental sem fins lucrativos, o parceiro não possui concorrentes diretos, mas sim hospitais associados em pesquisa e tratamento da doença. Muito para além disso, identificam-se organizações privadas que, no sentido figurado, concorrem por pacientes. Como fator de observação, identifica-se os altos investimentos em educação e saúde por parte de outras organizações.

2. **Poder de negociação dos fornecedores:** por necessitar de produtos extremamente refinados e de alto valor agregado, o parceiro possui fornecedores com alto poder de barganha, uma vez que certos medicamentos e aparelhos de pesquisa não são ofertados de maneira massiva e possuem suas produções e precificações controladas por um pequeno grupo de empresas. Como fornecedores de dados, os hospitais também podem negar o fornecimento da matéria prima das atividades desenvolvidas pelo parceiro: os dados. E, por fim, pelas tecnologias criadas necessitarem de um extenso processo de licenciamento, exigências quanto ao custo da solução podem ser feitos por aqueles que viabilizaram o processo de sua criação.

3. **Poder de negociação dos clientes:** como detentores de seus dados, os pacientes podem negar o uso de seus dados ao parceiro, inviabilizando, assim, a sua atuação em pesquisa e no desenvolvimento de novas tecnologias. Em outra perspectiva, eles também podem recusar o uso das soluções desenvolvidas impactando, também, à sua aderência e popularidade.

4. **Ameaça de novos entrantes:** com os recentes altos investimentos da iniciativa privada no desenvolvimento de novas tecnologias e serviços na mesma fatia de mercado do parceiro, identifica-se um risco moderado de aparecimento de soluções mais adequadas visto a abrangência e disponibilidade de recursos quase que inesgotáveis de fundos financeiros interresados.

5. **Ameaça de produtos substitutos:** criação de terapias alternativas para a manipulação de genes que causam a tendência do câncer, inviabilizando, assim, a necessidade das soluções propostas pelo parceiro.

4.1.2. Análise SWOT

4.1.3. Planejamento Geral da Solução

a) quais os dados disponíveis (fonte e conteúdo - exemplo: dados da área de Compras da empresa descrevendo seus fornecedores)

b) qual a solução proposta (pode ser um resumo do texto da seção 2.2)

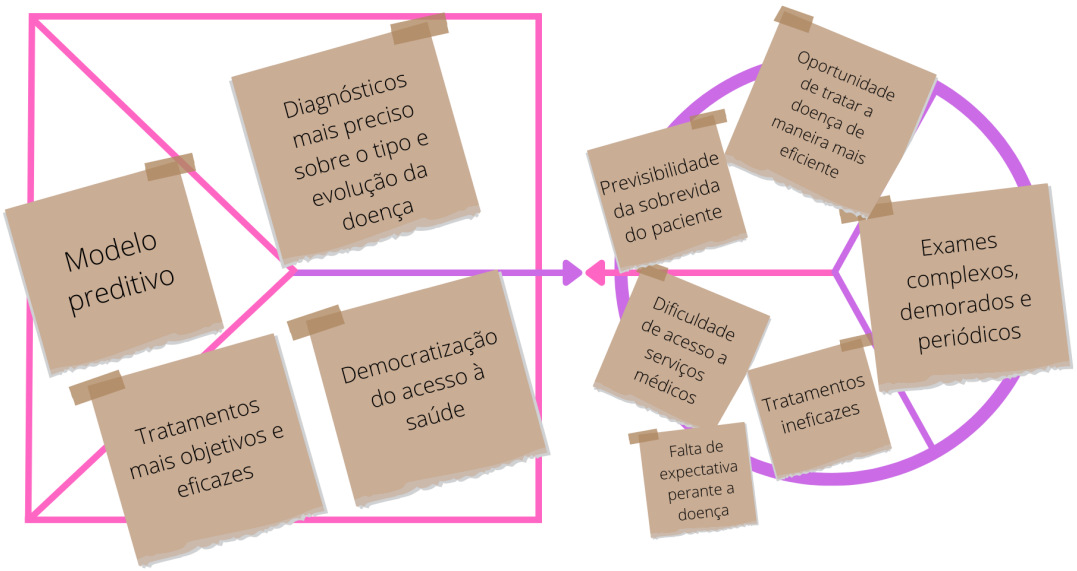
c) qual o tipo de tarefa (regressão ou classificação)

d) como a solução proposta deverá ser utilizada

e) quais os benefícios trazidos pela solução proposta

f) qual será o critério de sucesso e qual medida será utilizada para o avaliar

4.1.4. Value Proposition Canvas



4.1.5. Matriz de Riscos

Probabilidade	Ameaças					Oportunidades				
90%	-	-	**Ameaça 018:** nenhum padrão de predição detectado	**Ameaça 019:** análise SWOT incompleta	**Ameaça 020:** análise de mercado imprecisa	**Oportunidade 008:** refinar conhecimentos sobre desenvolvimento de algoritmos de aprendizado de máquina em geral	**Oportunidade 009:** refinar conhecimentos sobre design e experiência do usuário em modelos preditivos	**Oportunidade 010:** entendimento da fatia de mercado de atuação do cliente	**Oportunidade 011:** refinar conhecimentos sobre documentações de projetos de inteligência artificial	-
70%	-	-	-	**Ameaça 016:** falta de alinhamento com o cliente no que diz respeito ao significado de algumas especificidades médicas	**Ameaça 017:** definição de persona imprecisa no que diz respeito ao uso da solução, vontades, pensamentos e expectativas.	**Oportunidade 005** desenvolvimento de power-skills relacionadas ao setor da saúde	**Oportunidade 006:** aprender a lidar com imprevistos de maneira positiva	**Oportunidade 007:** assimilar conhecimentos relacionados a esse tipo de projeto	-	-
50%	**Ameaça 011:** atualização errônea no algoritmo da solução (mesclagem entre códigos incorreta que pode levar a digergências de interpretação)	**Ameaça 012:** atraso na entrega de possíveis funcionalidades extras solicitadas de última hora	**Ameaça 013:** integrante do projeto se ausentar por motivos pessoais	**Ameaça 014:** dados ausentes ou enviesados	**Ameaça 015:** distribuição inadequada do projeto	**Oportunidade 002:** cliente dar continuidade ao mínimo produto viável desenvolvido	**Oportunidade 003:** expandir a rede de contatos a partir das pessoas envolvidas no projeto	**Oportunidade 004:** adição do projeto ao portfólio	-	-
30%	**Ameaça 006:** back-log com informações desatualizadas e uso de metodologias ainda pouco exploradas	**Ameaça 007:** construção de um modelo preditivo não entendível	**Ameaça 008:** consumo de documentação ambígua	**Ameaça 009:** definição de histórias de usuário inapropriadas	**Ameaça 010:** interpretação incorreta ou enviesada dos dados	**Oportunidade 001:** captar novas propostas e projetos do cliente em razão do resultado do projeto	-	-	-	-
10%	**Ameaça 001:** definição de preferências de design e experiência do usuário para solução de inteligência artificial inadequadas	**Ameaça 002:** desenvolvimento de códigos de aprendizado de máquina não padronizados	**Ameaça 003:** construção de uma solução com baixa aderência do público por desconfiança	**Ameaça 004:** demora no processamento e operações de dados pelos servidores disponíveis	**Ameaça 005:** servidores de processamento de dados indisponíveis para uso	-	-	-	-	-
	Muito baixo	**Baixo**	**Moderado**	**Alto**	**Muito alto**	**Muito alto**	**Alto**	**Moderado**	**Baixo**	**Muito Baixo**

Impacto

4.1.6. Personas

Posicione aqui suas Personas (as que utilizam o modelo e as que são afetadas pelo modelo)

4.1.7. Jornadas do Usuário

4.2. Compreensão dos Dados

Os dados utilizados para a construção do modelo preditivo aqui disposto são oriundos, em sua maioria, dos prontuários eletrônicos de pacientes do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC/FM-USP). Esses dados representam mulheres diagnosticadas com câncer de mama em diferentes estágios e quatro subtipos: Luminal-A, Luminal-B, HER-2 e TNBC (triplo negativo).

Os bancos de dados disponibilizados encontram-se em diversos formatos, dos quais se sobressaem o CSV (valores separados por vírgula) e XLSX (planilha padrão do excel). O principal dos bancos, possui 61684 tuplas e 104 colunas, com registros de cerca de 3769 pacientes únicos. São, a princípio, para a solução desenvolvida, conteúdos relevantes dessa base:

- Escolaridade e nível de educação;
- Raça;
- Histórico de gravidez e quantidade de partos e abortos;
- Histórico de amamentação e tempo de amamentação;
- Menarca;
- Índice de massa corpórea;
- Tempo de reposição hormonal;
- Histórico de câncer na família; e
- Tempo de sobrevida calculado.

Devido a natureza dispersa dos dados, por meio dos identificadores únicos de cada paciente, visando a construção de um fluxo de trabalho mais sólido e consistente, uma mesclagem viabilizada a partir de soluções automáticas de limpeza e tratamento de dados se faz necessária.

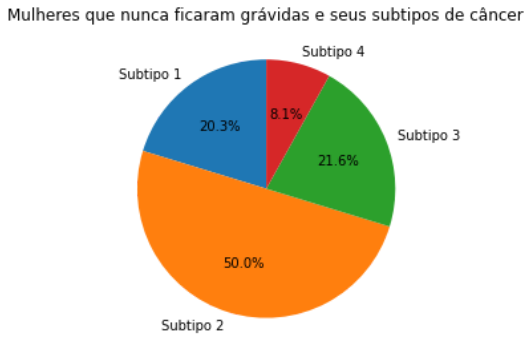
É entendível, também, o caráter enviesado das informações adquiridas, uma vez que o Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC/FM-USP) atua como agente médico-hospitalar de nível terciário de complexidade, recebendo pacientes que necessitam de tratamentos e terapias especializadas.

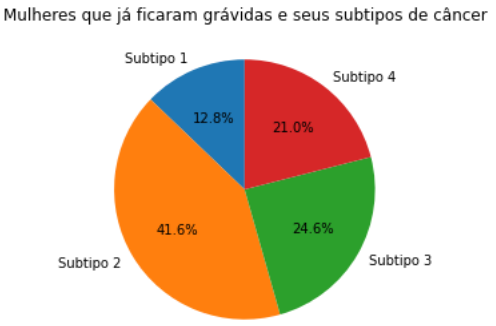
Como os dados fornecidos possuem problemas quanto a sua qualidade, cobertura e diversidade, tomam-se alguns riscos e contingências quanto a construção da solução, incluindo o problema de sobreajuste, que emerge-se como sendo a baixa capacidade de generalização do algoritmo devido a um super ajuste ao conjunto de dados de treinamento por falta ou excesso de informações. Na tentativa de atenuar esses problemas, o cruzamento de dados e replicação de informações foi realizado. Todavia, ainda não é possível determinar o grau de efetividade dessa ação.

Tratando da existência de subconjuntos de dados, uma vez que o tamanho original da base de dados impossibilita a sua utilização completa em todas as etapas da definição do modelo, evidenciam-se alguns parâmetros principais, por ordem de prioridade, quais utilizados para uma análise de primeira instância:

1. Subtipo;
2. Estadio;
3. Histórico de reposição hormonal;
4. Histórico de amamentação;
5. Histórico de gravidez; e
6. Resultados de hemogramas.

Como estudos iniciais, destacam-se os seguintes gráficos obtidos a partir de um algoritmo de construção estatística qualquer:





Por fim, após o entendimento da solução e seu objetivo, define-se o atributo alvo (saída esperada) do modelo preditivo como sendo o índice de sobrevida geral do paciente, de natureza binária, inicialmente dividido em baixo e alto, podendo, num futuro próximo, ser atualizado para gradientes escalares de quartis, possibilitando uma melhor acurácia e resposta.

4.3. Preparação dos Dados

Descreva as etapas realizadas para definir os dados e os atributos descritivos dos dados (“features”) a serem utilizados. Essa descrição deve ser feita de modo a garantir uma futura reprodução do processo por outras pessoas, e deve conter:

a) Descrição de quaisquer manipulações necessárias nos registros e suas respectivas features.

b) Se aplicável, como deve ser feita a agregação de registros e/ou derivação de novos atributos.

c) Se aplicável, como devem ser removidos ou substituídos valores ausentes/em branco.

d) Identificação das features selecionadas, com descrição dos motivos de seleção.

Não deixe de usar tabelas e gráficos de visualização de dados para melhor ilustrar suas descrições.

IMPORTANTE: Crie tópicos utilizando a formatação “Heading 3” (ou menor) para que o Google Docs identifique e atualize o Sumário (é necessário apertar o botão Refresh no Sumário para ele coletar as atualizações)

4.4. Modelagem

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

4.5. Avaliação

Nesta seção, 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. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

4.6 Comparação de Modelos

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

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

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.