NOME DO PROJETO
Faculdade de Medicina da USP

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
11/08/2022	Luísa Vitória Leite Silva, Moises Caze e Bianca Casemiro Lima	1.0	Objetivos e justificativa e preenchimento da seção 4

Sumário

1. Introdução

Apresente de forma sucinta o parceiro de negócio, seu porte, local, área de atuação e posicionamento no mercado. Maiores detalhes deverão ser descritos na seção 4

Descreva resumidamente o problema a ser resolvido (sem ainda mencionar a solução).

Caso utilize citações ao longo desse documento, consulte a norma ABNT NBR 10520. Sugerimos o uso do sistema autor-data para citações.

2. Objetivos e Justificativa

2.1. Objetivos

Os principais objetivos do parceiro de negócio consistem em fazer uma análise dos dados da paciente de câncer de mama e a partir dessa análise conseguir receber uma resposta sobre a taxa do tempo de sobrevida, mas não só isso como também um direcionamento sobre o agendamento de consultas futuras levando em conta o status da paciente e os possíveis desdobramentos de seu tratamento.

2.2. Justificativa

Levando em conta os objetivos citados no item anterior, nossa solução proposta é a criação de uma Inteligência Artificial que consiga analisar os dados da paciente, e a partir disso faça uma análise preditiva sobre a taxa do tempo de sobrevida e também com base nos dados principais e estruturados consiga direcionar o médico sobre os possíveis agendamentos de consulta.

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

1) A ameaça de novos entrantes é baixa, pois se trata de uma instituição pública, portanto novos hospitais não seriam uma ameaça ao Hospital da Usp. Entretanto, quando falamos de hospitais particulares a história é diferente, pois estes competem entre si pelos melhores números para dessa maneira receber reconhecimento.

2) A ameaça de novos produtos é baixa. Levando em conta que se trata de um hospital, os serviços realizados por essa instituição são considerados de extrema importância e um direito básico para a sociedade segundo o Art.196 da nossa Constituição Federal. Portanto, a ameaça

3) Poder de barganha dos compradores, em razão do hospital atender uma grande parcela da sociedade consideramos esse público pulverizado, por esse motivo há um baixo poder de barganha dos compradores.

4) Poder de barganha dos fornecedores - o poder de barganha dos fornecedores é alto, pois as instituições responsáveis por fornecer as máquinas do hospital tem uma alto poder nesse sentido, porém isso não acontece quando falamos da indústria farmacêutica que tem um poder mais baixo por ter uma concorrência maior.
pulverizado ou concentrado

5) Rivalidade entre competidores - o preço dos serviços oferecidos de certa forma são tabelados e são custeados pelo governo, não só isso mas como é um serviço público tal força diminui ainda mais,portanto a rivalidade entre competidores é baixa. Levando em conta a excelência do hospital da USP percebemos que ele está presente entre os 10 melhores hospitais do Brasil.

4.1.2. Análise SWOT

4.1.3. Planejamento Geral da Solução

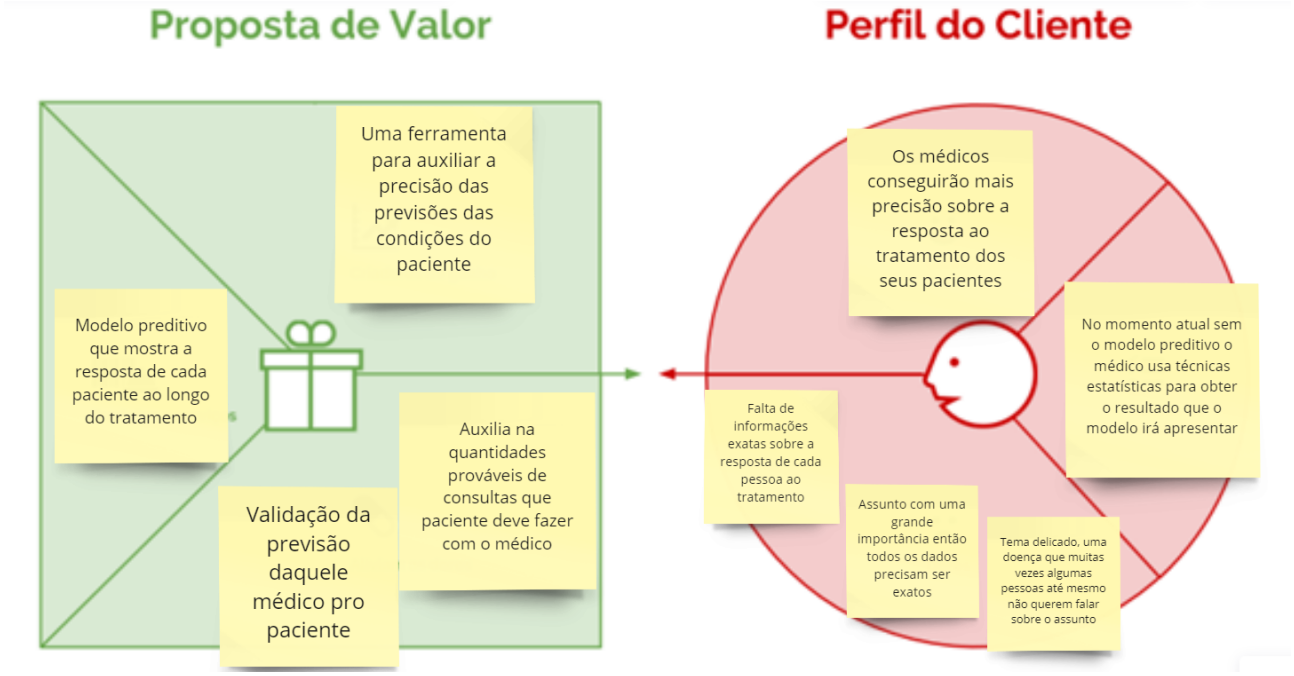
Uma pesquisa da CNN Brasil revelou que, segundo a diretoria da OMS, são quase 600 mil pessoas todos os anos que desenvolvem quadros de câncer e, dentre eles, o câncer de mama está entre os mais recorrentes. A Faculdade de Medicina da USP nos apresentou dados de prontuários médicos de pacientes que têm ou tiveram câncer de mama e utilizaram a rede pública de saúde desde 2008. Os *stakeholders* estão em busca de resultados preditivos, a partir da análise dos dados, sobre o tempo de sobrevida estimado em casos individuais.

Baseada em mais de 60 mil linhas de registros de cerca de 3 mil pacientes diagnosticados, nossa solução usa padrões identificados para estimar tempo de sobrevida, variabilidade da evolução do câncer de mama, da sua resposta aos tratamentos convencionais e marcadores que possam impactar a qualidade de vida, bem como a possibilidade de deixar sequelas. Nossa solução criará um score para classificar o tempo de sobrevida como alto ou baixo.

Por parte do Inteli, será desenvolvida uma inteligência artificial que classifica a sobrevida. Com isso, o trabalho dos médicos será facilitado, pois a estimativa de tempo de sobrevida dá ao médico assistência para administrar o tempo entre as consultas para acompanhamento. Com isso, aos stakeholders, médicos e pacientes, serão garantidos eficiência, eficácia e alcance na descoberta e luta contra a doença.

A inteligência artificial terá como critério de sucesso uma avaliação de, pelo menos, 0,8 na curva ROC. Desta forma, tornando claro o intervalo entre as consultas, o tempo será otimizado, trazendo maior número e qualidade para as consultas.

4.1.4. Value Proposition Canvas



4.1.5. Matriz de Riscos

Probabilidade	Ameaças					Oportunidades				
90%	Média	Média	Alta	Alta	Alta	Experiência em trabalhar com uma instituição renomada	Baixa	Baixa	Média	Média
70%	Baixa	Média	Média	Baixo conhecimento técnico das ferramentas	Modelo ter um baixo grau de confiança	Baixa	Experiência de trabalho em equipe	Média	Média	Alta
50%	Baixa	Baixa	Ter que coletar novos dados devido a baixa qualidade dos atuais	Alta	Faltou tempo para desenvolver	Baixa	Baixa	Identificar padrões nos dados para gerar insights relevantes	Alta	Alta
30%	Baixa	Baixa	Média	Pouco comprometimento da Medicina USP com a equipe	A equipe não ter bem definido quais são as métricas de avaliação do modelo	Conseguir novas variáveis que sejam mais relevantes para o treino do modelo	Conseguirmos prever outras variáveis além das esperadas	Média	Alta	Alta
10%	Baixa	Baixa	Falta de comprometimento do grupo em entregar as tarefas	Baixa	Ocorrer vazamento dos dados disponibilizados	Expansão do projeto para pacientes com outros tipos de câncer	Alta	Alta	Alta	Alta
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo
	Impacto									

4.1.6. Personas

NOME: Roger

IDADE: 57 anos

GÊNERO: Masculino

OCUPAÇÃO: Médico com especialização em Oncologia

"Minha profissão é minha paixão"

Considerações biográficas e comportamentais

- Trabalha com câncer há 15 anos
- Fluente em 3 idiomas
- Professor desde 2001
- Formado em Medicina pela USP em 1988

Dores/Motivações atuais com o problema:

- Quer proporcionar mais confiança aos pacientes
- Não tem uma ferramenta para validar sua provisão para o resultado do diagnóstico
- Ter uma base para saber quantas consultas deve fazer com cada paciente

Objetivos/necessidades específicas em relação ao problema:

- Otimizar o tempo de programar quantas consultas o paciente irá ter
- Consultas realmente necessárias com o paciente

NOME: Catarina

IDADE: 50

GÊNERO: Feminino

OCUPAÇÃO: Mãe e professora

"Catarina é mãe de dois filhos e recentemente foi diagnosticada com câncer de mama."

Considerações biográficas e comportamentais

- Fumante
- Prática pouco exercício físico
- Muito preocupada
- Passou da menopausa
- Desatenta consigo mesma

Dores/Motivações atuais com o problema:

- Deseja prolongar o seu tempo de vida
- Vive cansada e/ou debilitada
- Evitar idas desnecessárias ao médico

Objetivos/necessidades específicas em relação ao problema:

- Predições ao resultado do tratamento mais confiável
- Prolongar seu tempo restante de vida

4.1.7. Jornadas do Usuário

4.2. Compreensão dos Dados

1. O arquivo contendo a base de dados foi disponibilizado no formato xlsx. Para a leitura no pandas, o formato do arquivo foi transformado para csv (comma separated values). Estes dados são provenientes do banco de dados do Instituto do Câncer de São Paulo/Faculdade de Medicina da USP, com informações provenientes de pacientes mulheres diagnosticadas com câncer de mama, contendo 104 colunas e 61683 registros que totalizam 3769 pacientes únicos, dado que um paciente possui múltiplos registros. As colunas se referem a variáveis como subtipo do câncer, escolaridade, histórico de gravidez, histórico de consumo de álcool e tabaco, entre outros.

   1.1. Foram disponibilizadas três bases de dados. Uma delas é a descrita anteriormente, que possui dados sobre as pacientes e sobre seu quadro de saúde, a segunda, por sua vez, possui informações sobre altura, peso e a data em que essa coleta foi feita. Por fim, a terceira possui as datas das consultas das pacientes. Todas essas tabelas possuem o atributo "record_id", identificador das pacientes, possibilitando a mesclagem das tabelas por meio desta chave primária.

   1.2. Em relação aos dados dispostos, é possível observar grande presença de dados nulos, na maioria das colunas. Isso impossibilita algumas análises e torna o processamento dos dados mais trabalhoso devido a necessidade de lidar com esses valores, além disso, gera o risco de não possuirmos uma quantidade de dados significativa para o treino do modelo preditivo ou de um treino enviesado, devido a um tratamento inadequado desses valores.

   Os dados coletados representam pacientes que acessam a rede pública de saúde. Temos muitas variáveis, um risco pode ser utilizar variáveis inadequadas para treino do modelo

   1.3. Em relação à quantidade de registros, iremos fazer as análises baseadas em toda a base de dados, pois existem 3769 pacientes, o que não é uma quantidade tão grande que impossibilite a utilização do conjunto completo.Além disso, como possuímos muitas variáveis, decidimos inicialmente filtrá-las com base nos atributos que os especialistas da área

disseram que mais são relevantes para o problema que está sendo abordado. Algumas delas são: subtipos e estádios do câncer, histórico de gravidez e menstruação e consumo de álcool e tabaco.Ademais, caso descobrirmos algo considerável a respeito da eficiência das variáveis ao longo de nossas análises, iremos alterar o conjunto utilizado para nossas análises.

1.4. Para garantir a confidencialidade das pacientes, os dados foram anonimizados. Além disso, os dados disponibilizados são confidenciais e para uso único e exclusivo no desenvolvimento do projeto, portanto, não podem ser divulgados para o público.

2. Nossos processos da feature engineering se deram, principalmente, no preenchimento de dados nulos das pacientes baseado em algum outro registro em que os mesmos estivessem declarados, em caso de algum dado que se manteve fixo como a escolaridade, por exemplo, ou em caso de dados contínuos variáveis, onde um dos casos onde foi o do índice de massa corporal(imc ou bmi), onde calculamos a mediana entre os valores e usamos o resultado para calcular quando necessário. No processo de análises de correlações, buscamos identificar algumas tendências dentre as variáveis, então testamos algumas:

- A relação de **grau de escolaridade** das pacientes de acordo com seu **tempo de sobrevida**, pois imaginamos que o nível de escolaridade pode, por vezes, afetar a qualidade de vida da paciente, segue o gráfico no anexo I;
- A relação entre o **índice de massa corporal**(imc ou bmi) com a **taxa de sobrevida**, visando verificar se os extremos na escala do imc afetam ou não o tempo de vida da paciente, segue o gráfico no anexo II;

- Proporção para cada **subtipo** de acordo com o **tempo de sobrevida**, buscando identificar a proporção de tempo de sobrevida por subtipo do câncer(gravidade), segue o gráfico no anexo III.

3. De acordo com o Instituto nacional do câncer, o tempo de sobrevida (overall survival) é definido como:

"The length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive.

In a clinical trial, measuring the overall survival is one way to see how well a new treatment works. Also called OS".

- A variável target é "output_os";
- Ela possui dois valores possíveis (High OS e Low OS), portanto, tem natureza binária e faz com que tenhamos uma tarefa de classificação;
- Não possuímos registros em que ela é nula;
- Essa variável é bem balanceada, está em uma proporção quase de 50/50;
- Gráfico da proporção de possíveis valores da variável target no anexo IV.

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

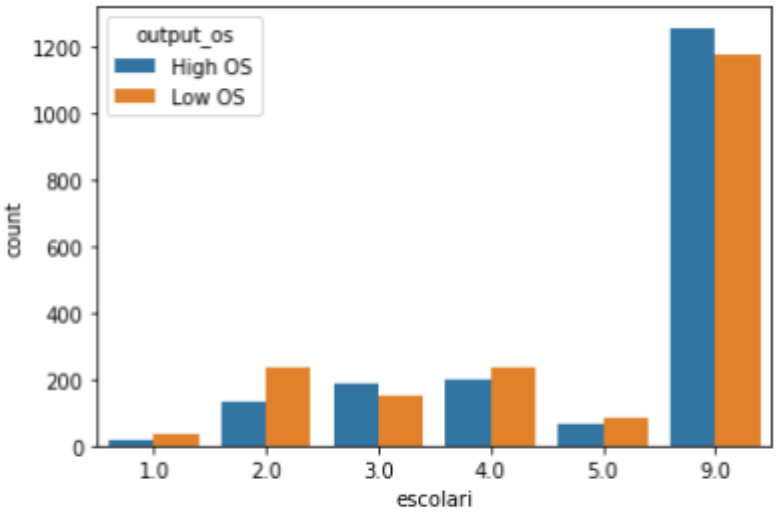
SERRANO, Layane; ROCHA, Lucas. **Brasil tem quase 600 mil novos casos de câncer por ano, diz diretora da OMS**. CNN Brasil, [*S. l.*], p. 1, 4 fev. 2022. Disponível em: https://www.cnnbrasil.com.br/saude/brasil-tem-quase-600-mil-novos-casos-de-cancer-por-ano-diz-diretora-da-oms/. Acesso em: 11 ago. 2022.

KLUYVER , Thomas; MCKINNEY, Wes. **Pandas: powerful Python data analysis toolkit**. [*S. l.*], 23 jun. 2022. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 12 ago. 2022.

NUMPY COMMUNITY. **NumPy User Guide**. [*S. l.*], 22 jun. 2022. Disponível em: https://numpy.org/doc/1.23/. Acesso em: 12 ago. 2022.

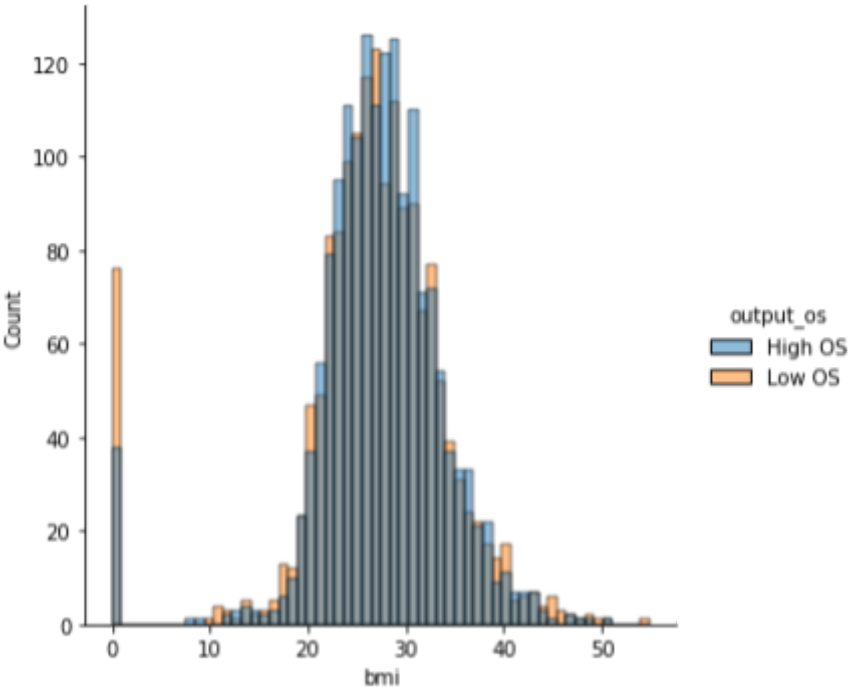
Anexos

Anexo I:



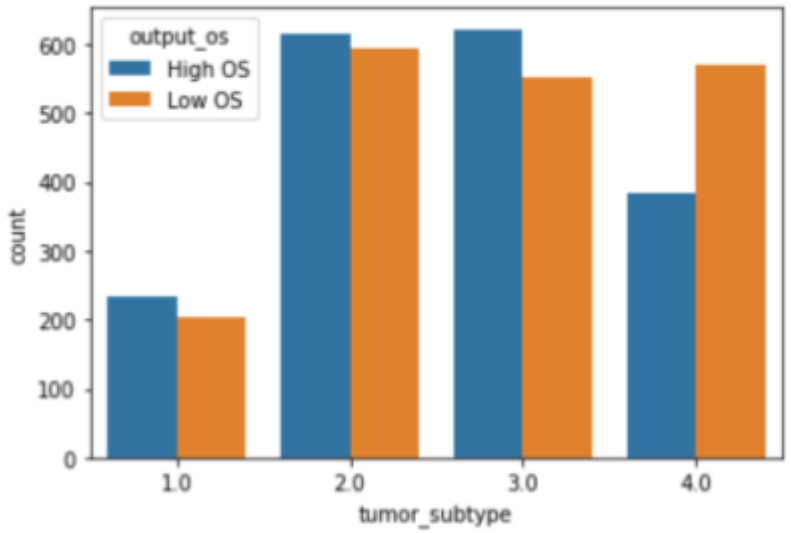
Relação da taxa de sobrevida em função da escolaridade

Anexo II:



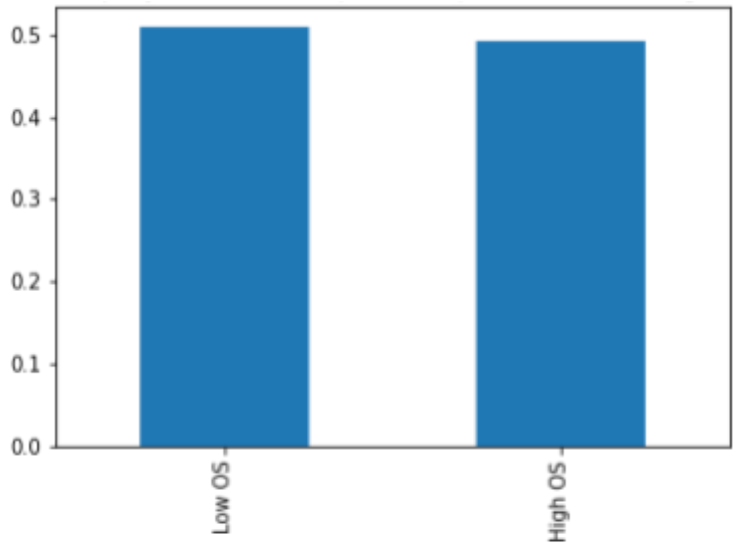
Relação da taxa de sobrevida em função do IMC(índice de massa corporal).

Anexo III:



Relação do tempo de sobrevida em função do subtipo do tumor

Anexo IV:



Relação de possíveis valores para a variável target.