Modelo de Predição Usp Medicina

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão (Sprint + Versão)	Resumo da atividade
01/08/2022	Dayllan Alho	1.1	Criação do documento
11/08/2022	Gabriel Nhoncanse	1.2	Inseri os modelos, que já estavam feitos, no documento, além de uma breve introdução de cada tópico.
11/08/2022	Jordan Andrade	1.3	Adição da introdução do documento revisão do documento com algumas alterações ortográficas
12/08/2022	Jordan Andrade	1.4	Revisão de alguns conceitos com base no encontro com o cliente
28/08/2022	Jordan Andrade , Dayllan e Henri	1.5	Preenchimento dos tópicos 4.3.1 ao tópico 4.3.14 acréscimo da jornada do usuário
09/08/2022	Henri Harari	1.6	Preenchimento do tópico 4.4
11/09/2022	Jordan Andrade / Oliver	1.7	Preenchimento do tópico 4.5
21/09/2022	Gabriel Nhoncanse / Henri Harari	1.8	Preenchimento do tópico 4.4 / 4.5
25/09/2022	Dayllan	1.9	Acréscimo de informação nos tópicos 4.4 / 4.5
25/09/2022	Gabriel Nhoncanse	2.0	Acréscimo de informação no tópico 4.4

Sumário

1. Introdução

O projeto apresentado tem como parceiro de negócios a Faculdade de Medicina da Universidade de São Paulo/Instituto do Câncer do Estado de São Paulo, instituição governamental voltada para atendimento de serviços médicos às comunidades assistidas pelo Sistema Único de Saúde(SUS). O hospital se encontra no endereço Av. Dr. Arnaldo, 251 - Cerqueira César - CEP 01246 903 - São Paulo - SP, e atua no tratamento e profilaxia de patologias humanas e prestação de serviços à comunidade, relacionadas à medicina, fisioterapia, fonoaudiologia e terapia ocupacional, dentro dos mais elevados preceitos éticos e morais. Atualmente é reconhecido como um dos hospitais de referência no combate ao câncer e reconhecido como o melhor hospital público do Brasil pelo World 's Best Hospitals 2022.

2. Objetivos e Justificativa

2.1. Objetivos

O parceiro tem como principal objetivo a criação de um sistema preditivo usando inteligência artificial com a finalidade de obter predições sobre a sobrevida da paciente, para assim conseguir definir o tratamento mais assertivo e menos nocivo com base nos resultados obtidos.

2.2. Proposta de Solução

O problema a ser resolvido pelo time é analisar a dificuldade de tratar o câncer de mama, devido às grandes variações de resultados em relação aos tratamentos convencionais. A partir dos dados disponibilizados podemos observar de forma um pouco mais aprofundada como a USP Medicina organiza e orquestra sua base de dados e o quais os dados indispensáveis para uma análise preditiva.

A solução proposta pelo grupo é um modelo preditivo que a partir da análise de dados clínicos definirá a situação entre normal e incomum. Isto é, será um prognóstico da condição da evolução do câncer de mama, pois a avaliação e veredito final serão ditados pelo profissional da área, que a partir de casos incomuns poderá dar seguimento com os pacientes.

O projeto executará a tarefa de classificação a partir de dados informados pelo médico durante o exame, para que a partir desses dados seja possível prever a evolução da variabilidade do câncer de mama e classificar o risco de vida do paciente. A solução poderá ser utilizada por meio de uma interface onde o profissional da saúde colocará as informações centrais do paciente e a IA analisará a evolução dos dados e do prognóstico.

Os benefícios do modelo de predição são muitos devido ao trabalho da inteligência artificial, como o ganho de tempo dos profissionais da área da saúde, que pode ser destinado ao tratamento de outros novos pacientes. Outro benefício é a precisão do prognóstico, que pode aumentar muito o número de tratamentos e de vidas salvas. Além disso, o encaminhamento de pacientes de acordo com o estágio cancerígeno e a possibilidade de evolução da doença com base na predição pode reduzir a fila de espera para o tratamento,uma vez que no cenário atual todos os pacientes devem comparecer ao hospital na mesma frequência visto que não se sabe qual será a reação do enfermo ao tratamento, o que o ocupa as vagas e sobrecarrega o hospital.

O critério de sucesso será após a conclusão do médico a partir dos dados obtidos, e então, após analisados pela IA, a forma a qual será utilizada para avaliar. Ao concluir a etapa da doença que o paciente se encontra, assim, será obtida a classificação de risco do paciente.

2.3. Justificativa

A proposta de solução é um modelo preditivo que auxilie o médico a escolher o melhor tratamento para a paciente, tendo em vista dados sobre a variabilidade do câncer de mama e sobre o próprio paciente. Um diferencial do nosso modelo é que ele funciona completamente por meio de inteligência artificial, a qual será treinada por meio de dados de casos passados e conseguirá dar uma predição mais precisa.

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 Collaboratory)

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

Nesse mesmo setor, o Instituto do Câncer de São Paulo não possui necessariamente concorrentes, pois não disputa mercado com outros players, porém, existem outras empresas que trabalham no mesmo setor, como por exemplo: AC Camargo, hospital oncológico especializado no diagnóstico, tratamento e pesquisa de câncer em humanos, Instituto nacional do Câncer, órgão auxiliar do Ministério da Saúde no desenvolvimento e coordenação das ações integradas para a prevenção e o controle do câncer no Brasil, e a FEMAMA, associação civil, sem fins econômicos, que busca ampliar o acesso ao diagnóstico e tratamento do câncer de mama para todas as pacientes, buscando reduzir os índices de mortalidade pela doença no Brasil.

Além disso, hoje o hospital se mantém ativo através de recursos públicos vindos majoritariamente do Ministério da Saúde, porém, podem vir de outras fontes governamentais, assim oferecendo um serviço gratuito aos pacientes.

Tendo em vista o mercado que o hospital está inserido, conclui-se que ele está constantemente buscando se aperfeiçoar e inovar, contando com investimentos públicos e privados à procura de inovações, como equipamentos novos, tratamentos mais eficazes e menos nocivos etc.

Em relação às **5 forças de porter**, concluímos que:

Rivalidade entre os concorrentes: Por ser uma instituição governamental sem fins lucrativos, o Instituto do Câncer de São Paulo não possui concorrentes de fato e sim hospitais parceiros em pesquisa e tratamento do câncer. Todavia, existem outras instituições como o Instituto Nacional do Câncer e diversos hospitais privados que realizam o tratamento de câncer e "disputam" pacientes.

Poder de negociação dos fornecedores: Por necessitar de produtos extremamente refinados e de alto valor agregado, o Instituto do Câncer de São Paulo possui fornecedores com alto poder de negociação, visto que certos medicamentos e aparelhos de pesquisa não possuem grande oferta no mercado e possuem sua produção e precificação controladas por um pequeno grupo de empresas.

Ameaça de entrada de novos concorrentes: Não há ameaça de entrada de novos concorrentes, pois as instituições da mesma área de trabalho da USP medicina não se passam

por concorrentes, e sim por parceiros. Além disso, a USP é ponto de referência na área de saúde e qualquer empresa que surgisse não conseguiria a curto prazo ser um concorrente à altura.

Ameaça de produtos substitutos: Pelo tratamento oncológico ser uma área que ainda possui muitos mistérios para a medicina e não possuir uma exatidão no tratamento e evolução dessa patologia, os riscos de surgir algum produto que substitua os métodos de tratamento convencional do câncer são praticamente nulos a curto prazo.

Poder de negociação dos clientes: Por ser um órgão público que atua predominantemente com pessoas de baixo poder aquisitivo, os clientes do Instituto do Câncer de São Paulo possuem baixo poder de negociação, uma vez que depende do Sistema Único de Saúde (SUS) para realizar o tratamento e não possuem condições de arcar com os custos de um hospital privado

4.1.2. Análise SWOT

A meta da análise SWOT é facilitar na identificação de características da empresa parceira (USP Medcina) e do mercado em que ela se encontra, assim nos ajudando no desenvolvimento do projeto. Além disso, ela facilita a potencialização de suas forças, mitigação de suas fraquezas e minimização de erros, procurando oportunidades para melhorar seus produtos ou elaborar novos protótipos. Diante disso, foi montada uma análise SWOT com base nas características do parceiro de negócios, que pode ser visualizada na imagem abaixo:

Imagem 1: Análise SWOT da USP Medicina

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

É uma ferramenta desenvolvida com a meta de explorar mais profundamente o cliente e a relação dele com o nosso produto por meio de uma análise das suas dores e como o nosso software irá saná-las.Diante disso, foi elaborado um modelo de Value Proposition Canvas com base nas dores do parceiro de negócio e a solução pensada pelo time de desenvolvimento

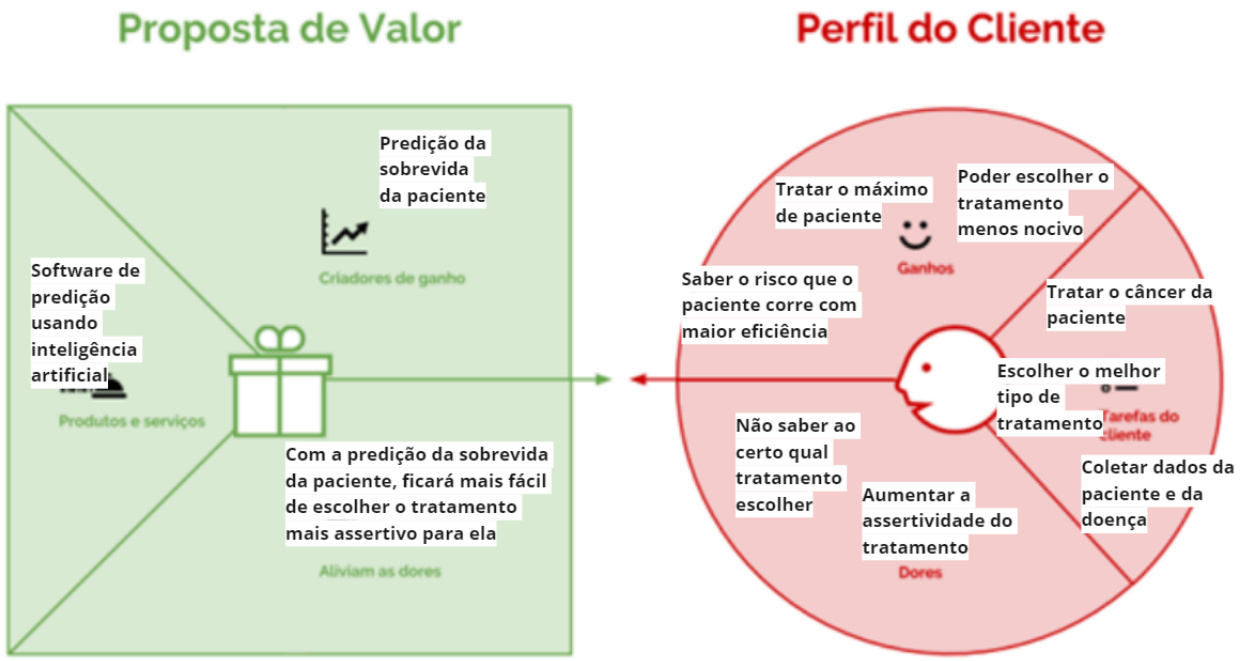


Imagem 2: Value Proposition Canvas

4.1.5. Matriz de Riscos

Também chamada de matriz de probabilidade e impacto, a matriz mapeia os riscos do projeto, sejam eles tanto riscos de ameaças quanto de oportunidades. Por ser uma ferramenta útil para gerenciar os riscos operacionais existentes em um projeto, foi elaborado uma Matriz de Riscos com base na proposta de solução elaborada pelo time de desenvolvimento, que pode ser visualizada na imagem a seguir:

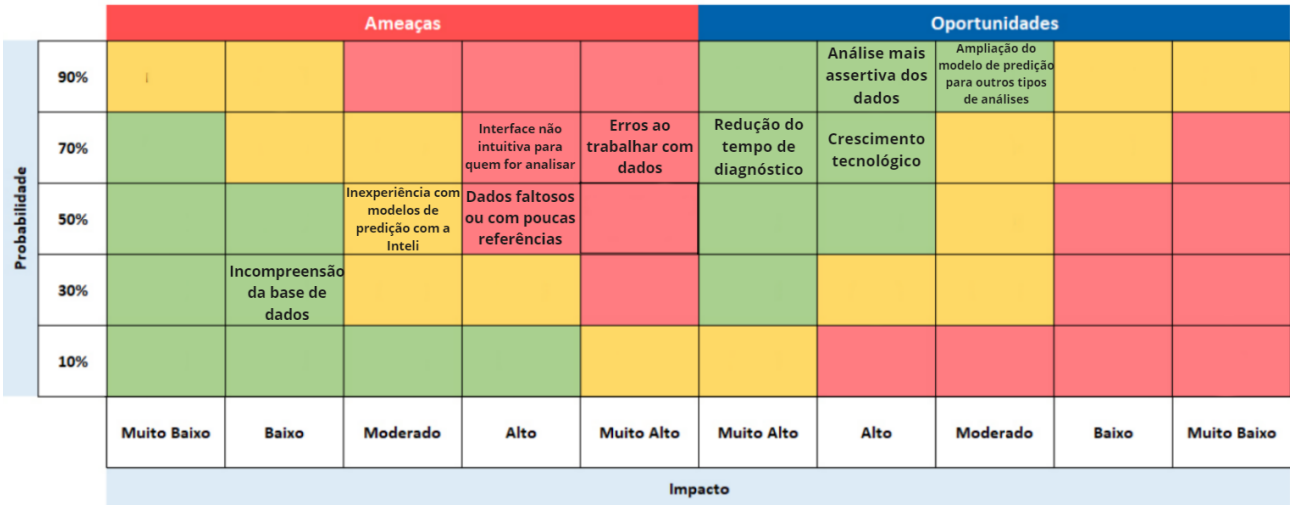


Imagem 3: Matriz de riscos

4.1.6. Personas

Persona 1 - Médico

Nome: James Wilson da Silva
Idade: 37 anos
Gênero: Masculino
Ocupação: Oncologista

“Um médico nunca desiste de seus pacientes”

Imagem 4: persona 1-James Wilson da Silva

Personalidade: Boa comunicação, apesar de ser introvertido; Gosta de passar seu tempo livre com a família; É uma pessoa persistente e está disposto a tudo para salvar seus pacientes.
Biografia: Pai de 3 filhos, James é chefe de Oncologia do Hospital Universitário da USP e lidera pesquisas sobre o câncer de mama. Durante sua adolescência, sua mãe foi diagnosticada com câncer de mama e, infelizmente, não sobreviveu. Esse ocorrido o motivou para se tornar um médico especializado em câncer de mama.

Motivações para usar o novo sistema: A perda de sua mãe pelo câncer de mama, fazendo com que ele procure um tratamento mais assertivo e eficaz.
Dores com o atual sistema:
1- Não saber ao certo qual tratamento seria melhor para determinada paciente;
2- Tratamentos às vezes mais agressivos do que o necessário.
Objetivos com o novo sistema: Garantir um prognóstico preciso e menos agressivo possível aos pacientes;

Persona 1 - Paciente

Nome: Lisa Andressa dos Santos
Idade: 45 anos
Gênero: Feminino
Ocupação: Professora do Fundamental

“Se há um impossível, meu criador não o conhece”

imagem 5: persona 2-Lisa Andressa dos Santos

Personalidade: Muito inteligente e paciente; Sonhadora e esperançosa; Mulher de fé; Mãe de 2 filhos; Adora crianças.
Biografia: Nascida e criada na Zona Leste de São Paulo, Lisa sempre foi apaixonada pela educação, paixão essa que a motivou a se tornar professora e, por meio do ensino e do afeto com seus alunos, criar uma sociedade melhor alterando as bases de ensino. Infelizmente, foi diagnosticada com câncer de mama há cerca de 2 anos, o que a levou a ser afastada do cargo de professora. Entretanto, Lisa é uma mulher sonhadora e cheia de fé, e acredita na progressão positiva do tratamento para poder voltar a dar aulas e ver seus filhos e alunos crescerem e terem um futuro brilhante.

Motivações para usar o novo sistema: Ver o crescimento dos filhos; Voltar a dar aula; Busca por um tratamento eficaz.
Objetivos com o novo sistema: Ser curada; Ter um retorno rápido sobre a situação da evolução do câncer; Conseguir um tratamento menos agressivo.

4.1.7. Jornadas do Usuário

A jornada do usuário é um mapa visual de todas as etapas do relacionamento da persona com o produto ou serviço oferecido por uma empresa.Ela descreve o passo a passo percorrido pelo usuário, detalhando todos os pontos de contato e interações do ponto de vista dele, seus sentimentos e sensações em cada fase, sendo possível a partir desse feedback da interação identificar pontos de melhoria e ajustes no produto. Dessa forma, foi realizada uma Jornada do Usuário com base na solução proposta pelo time de desenvolvimento para o problema acerca da variabilidade do câncer de mama, que pode ser visualizada a seguir:

Clique aqui caso queira uma melhor visualização.

Imagem 6: Jornada do usuário

4.2. Compreensão dos Dados

Tendo como princípio os dados fornecidos pelo Instituto do Câncer de São Paulo, coletados de prontuários eletrônicos de pacientes diagnosticados com câncer de mama em diferentes estágios, foram disponibilizados no formato csv e xml.

Com o recebimento de dados, foi realizado uma triagem da relevância das informações, com um aprofundamento no conhecimento das informações relativas ao câncer, e a posteriori levantadas possibilidades de teorias que a predição poderia realizar a partir de novas informações.

Ademais, com uma análise da base disponibilizada, foi possível observar uma certa instabilidade na base, ausência de dados relevantes e a confusão na identificação dos itens das colunas, colaboram em ampliar a dificuldade da predição.

- Última informação do paciente: status se está vivo ou morto, para evaluar a perspectiva de vida conforme determinados parâmetros;
- consumo de álcool: procurar relação com o progresso do câncer;
- consumo de tabaco: procurar relação com o progresso do câncer
- Possui histórico familiar: procurar relação hereditária do câncer
- Recidiva: verificar casos em que houve reincidência do câncer para determinado tratamento.
- Amamentou na primeira gestação? Validar a correlação da existência ou não do câncer pós amamentação.
- Por quanto tempo amamentou: Validar a correlação do câncer com o período de amamentação.

Idade da primeira menstruação: avaliar relações do câncer e períodos destrutivos. Os parâmetros foram obtidos a partir do arquivo csv “BDIPMamaV11 - DATA LABELS”.

Sendo os dados usados para fins estudantis do inteli, foi mapeado algumas colunas definidas como principais para obtenção de predições que trazem clareza para as perspectivas do médico e do paciente.

Portanto, utilizando de todos os artefatos de dados científicos, o algoritmo será direcionado a buscar um melhor mapeamento do câncer e seu progresso, conforme os parâmetros estabelecidos para calcular uma perspectiva de vida ou tratamento.

4.3. Preparação dos Dados

Com a finalidade de compreender melhor o banco de dados, o que eles representam e como se correlacionam com o projeto abordado, foi realizado pelo time de desenvolvimento uma feature engineering , na qual foram analisadas, selecionadas e tratadas todas as colunas que acredita-se possuírem alguma correlação com a variabilidade do câncer de mama. A seguir, é possível visualizar os dados selecionados bem como seu tratamento e importância:

4.3.1. Record ID (coluna A)

Essa coluna representa o número de identificação do paciente e aparece no banco de dados como um número cardinal que é exclusivo para cada paciente, que se repetia a cada linha conforme o paciente realizava uma nova consulta. Essa tabela foi selecionada para que pudesse ter o controle de quais atributos pertenciam a cada paciente e, para o tratamento dessa coluna, foram retiradas as linhas multiplicadas do mesmo ID e foi assumido como hipótese que a última informação do paciente é a mais importante (neste momento), sendo a essa a única linha que permaneceu.

4.3.2.Redcap_repeat_instrument(coluna B)

A coluna “redcap_repeat_instrument” possuía 3 dados diferentes(“registro de tumores”, “dados histopatológicos da mama” e “dados antropométricos”), sendo necessário transformar esses dados em 3 novas colunas de mesmo nome para que fosse trabalhado à parte cada feature. Entretanto, até o momento só foi utilizado a feature “dados antropométricos” para serem analisados quais exames a paciente realizou e compreender melhor o estágio atual do câncer dessa pessoa.

4.3.3.Dob (coluna D)/Date_last_fu(coluna BO)

A coluna “dob” (date of birthday) representa a data de nascimento da paciente e aparece no banco de dados no formato de data (dd/mm/aaaa), enquanto a coluna Datre_last_fu representa a data da última informação do paciente, e também se encontra no formato de data. Em uma hipótese levantada sobre os dados analisados, acredita-se que a idade da paciente possui influência sobre a resposta do organismo ao câncer, e portanto foi deduzido que a idade da paciente seria a diferença da última vez que foi visto e sua data de nascimento. Dessa forma, essas duas colunas foram devidamente tratadas(retirada de NaN e linhas vazias) e geraram uma nova coluna “idade” com base no cálculo citado.

4.3.4. Menarche(coluna M)/Period(coluna N)

A coluna “Menarche” representa a idade em que a paciente teve sua primeira menstruação, e aparece no banco de dados como um número cardinal que representa essa idade, enquanto a coluna Period indica o estado da menopausa da paciente, e aparece no

banco de dados como 0(pós-menopausa) e 1(pré-menopasua).
Em uma hipótese levantada sobre os dados analisados e conversas com o doutor Roger, acredita-se que o tempo de exposição hormonal da paciente(tempo do período fértil) tenha influência sobre a progressão do câncer. Dessa forma , ambas as colunas foram devidamente tratadas (retirada de linhas vazias e transformação de NaN para 0.0) e posteriormente serão usadas para realizar o cálculo desse tempo de exposição, utilizando a diferença da última e a primeira menstruação.

4.3.5.BMI (coluna O)

A coluna "BMI" representa o Índice de Massa Corporal(IMC) da paciente, e aparece no banco de dados como um número cardinal que indica esse índice. Em uma hipótese levantada com base na análise dos dados e pesquisas sobre o assunto, acredita-se que quanto maior o IMC e consequentemente maiores as chances de problemas relacionados à obesidade, maior o agravamento do estado de saúde da paciente e maiores os riscos de progressão do câncer. Sendo assim, essa coluna foi devidamente tratada(retirada de linhas vazias e transformação de NaN em 0.0) e a coluna permaneceu com os números cardinais indicando o IMC.

4.3.6.Antec_fam_cancer_mama (coluna Z)/Familial_degree____1(coluna AH)/Familial_degree____2(coluna AI)/ Familial_degree____3(coluna AJ)

A coluna "Antec_fam_cancer_mama" representa o histórico familiar de câncer de mama na família, e aparece no banco de dados no formato de string, sendo elas "sim" e "não", enquanto as colunas "familial_degree____1/2/3" representam o grau de parentesco da paciente com o parente vítima de câncer, sendo representado por 0(não) e 1(sim) caso aquela coluna represente o grau de parentesco de ambos . Em uma hipótese levantada com base na análise dos dados, pesquisas sobre o assunto e conversas com o doutor Roger , acredita-se que um histórico familiar de câncer pode indicar uma anomalia genética que predispõe a ocorrência da doença, sendo o grau de parentesco um fator agravante dessa situação (quanto mais próximo, maior a chance de ocorrer). Dessa forma, todas as colunas foram devidamente tratadas(retirada de colunas vazias e transformação de NaN em 0.0) e a coluna "Antec_fam_cancer_mama" foi trasnformada em outras 2 colunas (Antec_fam_cancer_mama_Não e Antec_fam_cancer_mama_Sim), ambas indicando valores de 0(não) e 1(sim).

4.3.7.Tobaco(coluna AA)

A coluna "Tobaco" indica se a paciente é usuária de cigarro, e aparece no banco de dados como 1 (nunca fumou),2(fumou no passado) , 3(fuma atualmente) e 99(não informado). Em uma hipótese levantada com base nos dados e pesquisas sobre o assunto, acredita-se que o consumo de cigarro reduz a eficácia do sistema imunológico e diminui sua ação contra células cancerígenas . Dessa forma, foi realizado o devido tratamento dos dados(retirada de linhas vazias e transformação de NaN em 0.0) e a coluna permaneceu nesse modelo de classificação de 1 a 4.

4.3.8.Alcohol(coluna AB)

A coluna “Alcohol” indica se a paciente é consumidora de bebidas alcóolicas, e aparece no banco de dados como 1(nunca bebeu), 2(bebeu no passado) , 3(bebe atualmente) e 99(não informado). Em uma hipótese levantada com base na análise dos dados e em pesquisas sobre o assunto, acredita-se que o consumo de álcool é um fator agravante na progressão do câncer, uma vez que corrompe o fígado(órgão vital para a regulação do metabolismo . Dessa forma, foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0) e a coluna permaneceu nesse modelo de classificação de a 1 a 4.

4.3.9.Primary_diganosis(coluna AX)

A coluna “Primary_diagnosis” representa o tipo histológico do tumor, e aparece na coluna como dados categóricos que vão de 1 ao 21, sendo cada um desses um tipo diferente de tumor. Em uma hipótese levantada com base na análise dos dados e pesquisas sobre o assunto, acredita-se que cada tipo tumoral tenha um impacto diferente sobre a progressão do câncer, sendo esse tumor mais agressivo ou não dependendo da sua classificação. Dessa forma,foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0) e os dados categóricos permaneceram nesse modelo de 1 a 21.

4.3.10.Benign(coluna AW)

A coluna “Benign” indica se o tumor analisado é benigno ou maligno, e aparece no banco de dados como 1(benigno) e 2(maligno).Em uma hipótese levantada com base em pesquisas e análise do banco de dados, acredita-se que tumores benignos possuem impactos irrelevantes ao corpo e à saúde do indivíduo, enquanto os tumores malignos são mais agressivos e requerem um maior cuidado médico. Dessa forma,foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0), e a classificação da coluna em 1 ou 2 se manteve.

4.3.11.Tumor_subtype(coluna BI)

A coluna “Tumor_subtype” representa o subtipo tumoral da paciente, sendo esse classificado de 1 a 4 no banco de dados..Em uma hipótese levantada com base em pesquisas e análise do banco de dados, acredita-se que que cada subtipo tenha um impacto diferente no organismo da paciente, sendo na maioria dos casos o tipo 4 mais agressivo ao corpo e o tipo 1 o menos agressivo. Dessa forma,foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0), e a classificação da coluna de 1 a 4 se manteve.

4.3.13.Follow_up_days(coluna BP)

A coluna “Follow_up_days” representa o tempo de seguimento do tratamento da paciente, desde o dia do diagnóstico até o dia em que se teve as últimas informações dela, e aparece no banco de dados como um número cardinal que indica esse tempo em dias . Essa coluna é de extrema importância para, a partir do tempo de vida das pacientes, identificar o comportamento da doença nos diversos casos e predizer como ela irá agir nas próximas

vítimas com base no padrão analisado. Dessa forma, foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0) e os números indicando quantos dias durou o tratamento se mantiveram.

4.3.14.Ultinfo(coluna BQ)

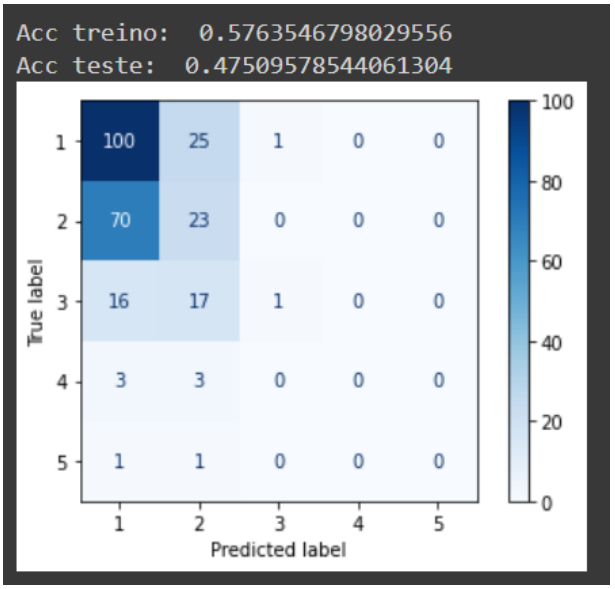
A coluna “Ultinfo” representa como a paciente se encontra desde a última vez que foi vista, e aparece na tabela como 1(vivo com câncer), 2(vivo SOE), 3(óbito por câncer) e 4(óbito por outras causas SOE). Essa coluna é de extrema importância para o time de desenvolvimento compreender a evolução da paciente diante do tratamento e a partir dos resultados da paciente predizer como futuras vítimas de câncer irão responder a esse mesmo tratamento com base no comportamento observado.Dessa forma, foi realizado o devido tratamento dos dados (retirada de linhas vazias e transformação de NaN em 0.0) e o modelo de classificação de 1 a 4 se manteve.

4.4. Modelagem

Durante o desenvolvimento da Sprint 4, decidimos por coletar alguns modelos de predição para melhor visualizar a dinâmica do treinamento e teste do algoritmo, para isso, utilizamos cinco grandes modelos, sendo eles: o KNN (KNowledge Now), a Árvore de Decisão, SVM, Regressão Linear e Regressão Logística. Desses modelos, percebemos que a regressão linear e a regressão logística não faziam muito sentido para nosso problema, pois é preferível utilizar um modelo de classificação ao invés de um modelo de regressão.

Assim, durante a modelagem selecionamos três modelos de treinamentos e testes que foram julgados mais relevantes para colocar em prática neste momento no projeto, onde é possível observar abaixo, uma pequena descrição deste modelo e quais foram os resultados encontrados

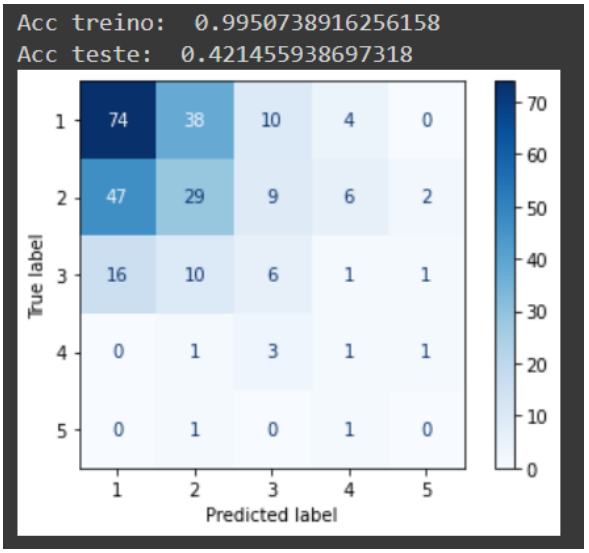
KNN (KNowledge Now): A ideia principal do KNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento. O algoritmo classifica basicamente g(X) atribuindo a ele o rótulo representado mais frequentemente dentre o número vizinhos mais próximos, no espaço onde se encontra as features selecionadas até então para o projeto. A principal vantagem que o grupo se conscientizou na escolha do KNN como um dos modelos testados foi: eficácia e simplicidade para obter uma fácil implementação em diversas situações. Assim, a acurácia de treino foi de 57%, e a de teste 47%. Abaixo pode ser vista a acurácia, representada pela sigla Acc, tanto de treino quanto a de teste.



Juntamente com a acurácia foi necessário observar a precisão, recall e f1-score do modelo, a fim de saber como a predição tem se comportado em relação aos falsos positivos.

	precision	recall	f1-score	support
1	0.54	0.59	0.56	126
2	0.37	0.31	0.34	93
3	0.21	0.18	0.19	34
4	0.08	0.17	0.11	6
5	0.00	0.00	0.00	2

Árvore de decisão: Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas, no nosso caso, com as escolhas das features. Permite que um indivíduo ou organização compare possíveis ações com base em seus custos, probabilidades e benefícios. Escolhemos esse método com o objetivo de criar um modelo que preveja o valor de uma variável de destino aprendendo regras de decisão simples inferidas a partir dos recursos de dados. A acurácia de treino após usarmos esse modelo foi de 42%.

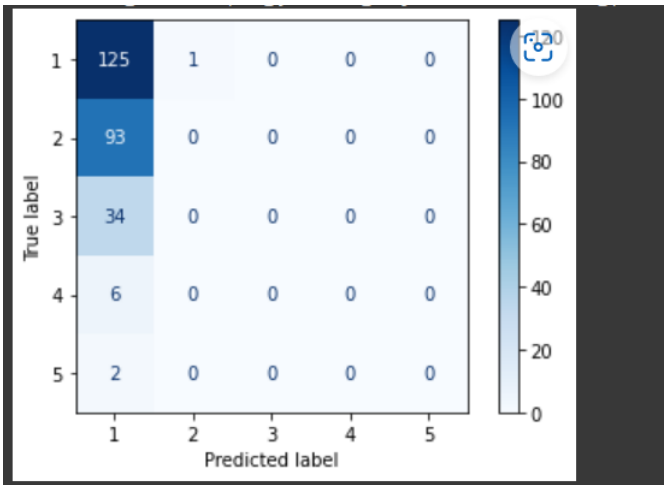


Juntamente com a acurácia foi necessário observar a precisão, recall e f1-score do modelo, a fim de saber como a predição tem se comportado em relação aos falsos positivos.

	precision	recall	f1-score	support
1	0.54	0.59	0.56	126
2	0.37	0.31	0.34	93
3	0.21	0.18	0.19	34
4	0.08	0.17	0.11	6
5	0.00	0.00	0.00	2

Regressão linear: A análise de regressão linear é usada para prever o valor de uma variável com base no valor de outra. A variável que deseja prever é chamada de variável dependente. A variável que é usada para prever o valor de outra variável é chamada de variável independente. Os dados foram submetidos ao processo de normalização padrão, que se baseia em limitar os dados entre 0 e 1. A utilidade desse processo se baseia em deixar os dados dimensionados da mesma forma o que resulta em uma melhor relação de proporcionalidade entre os valores das diferentes features. A porcentagem de de acurácia foi de 41%.

SVM: O SVM é um algoritmo que busca uma linha de separação entre duas classes distintas, analisando os dois pontos, um de cada grupo, mais próximos da outra classe. Isto é, o SVM escolhe a reta entre dois grupos que se distancia mais de cada um e, após descoberta essa reta, o programa conseguirá predizer a qual classe pertence um novo dado ao checar de qual lado da reta ele está. Com as variáveis (selecionadas previamente) já inseridas no modelo, conseguimos alcançar 47,8% de score com o kernel do tipo “poly”, que significa que a função reta traçada pelo modelo será um polinômio.



Acc treino: 0.5599343185550082
Acc teste: 0.4789272030651341

Hiperparâmetros

A) No knn, em um primeiro momento, foram utilizados os parâmetros

```
'n_neighbors':[1,2,33,5,13,7,20,3],

    'weights':['uniform','distance'],

    'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'],

    'leaf_size':[1,2,33,5,13,7,20,3,23]
```

,

obtendo o resultado

```
knn_best= randomized_search_knn.best_estimator_
knn_best
print('Acc treino:',knn_best.score(X_train,Y_train))
print('Acc teste:',knn_best.score(X_test,Y_test.squeeze()))
print('Revocação:',recall_score(Y_test,Y_par_knn,average=None))
print('Precisão:',precision_score(Y_test,Y_par_knn,average=None))
print('F1_score:',precision_score(Y_test,Y_par_knn,average=None))
```

```
Acc treino: 0.5763546798029556
Acc teste: 0.4521072796934866
Revocação: [0.81746032 0.16129032 0.         0.         0.        ]
Precisão: [0.49282297 0.28846154 0.         0.         0.        ]
F1_score: [0.49282297 0.28846154 0.         0.         0.        ]
```

,

com o auxílio do Grid Search e Randomized Search obtemos os melhores hiperparâmetros, que elevou nossa acurácia em 2%

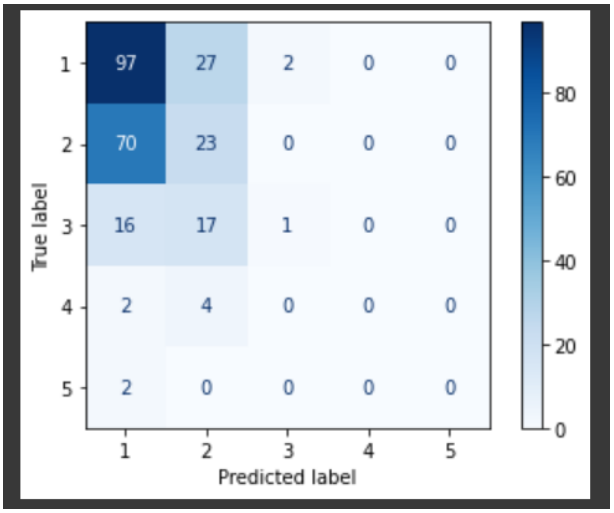
```
[38]
parametros_knn={
    'n_neighbors':[1,2,33,5,13,7,20,3],
    'weights':['uniform','distance'],
    'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'],
    'leaf_size':[1,2,33,5,13,7,20,3,23],
}
randomized_search_knn = RandomizedSearchCV(estimator = KNeighborsClassifier(),param_distributions=parametros_knn)
randomized_search_knn.fit(X_train,Y_train.squeeze())
Y_par_knn= randomized_search_knn.predict(X_test)
```

```
print(randomized_search_knn.best_score_)
print(randomized_search_knn.best_params_)
```

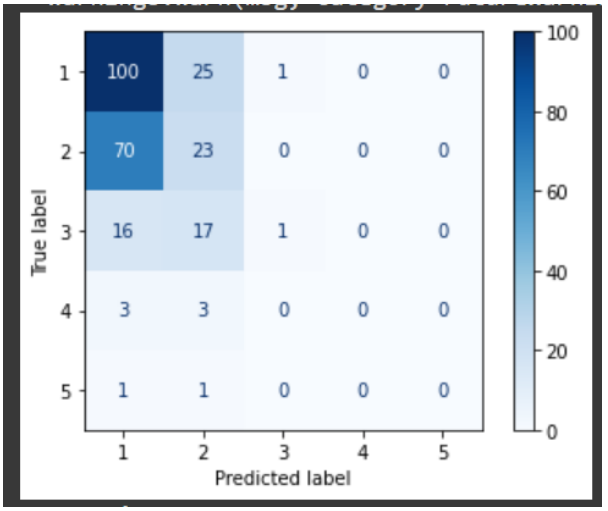
```
0.5614686356862213
{'weights': 'uniform', 'n_neighbors': 33, 'leaf_size': 2, 'algorithm': 'ball_tree'}
```

Sendo assim, os resultados comparados podem ser observados abaixo na matriz de confusão antes do hiperparâmetro e a matriz de confusão depois da implementação dos hiperparâmetros.

Com hiperparâmetros:



Sem hiperparâmetros:



utilizamos estes hiperparâmetros pois acreditamos que estes são os melhores encontrados para a aumentar a acurácia dos dados.

B)Na árvore de decisão, em um primeiro momento, foram utilizados os parâmetros

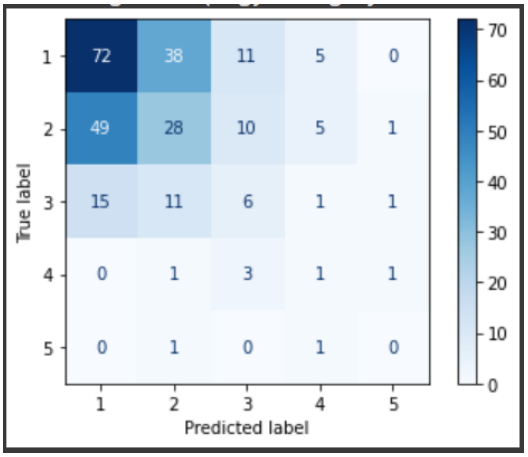
```
parametros_tree = {
    'criterion':['gini', 'entropy', 'log_loss'],
    'splitter':['best', 'random"'],
    'max_depth':range(2,16,2)
}
```

obtendo o resultado _____

com o auxílio do Grid Search e Rendomized Search obtemos os melhores hiperparâmetros, que elevou nossa acurácia em x%

Sendo assim, os resultados comparados podem ser observados abaixo na matriz de confusão antes do hiperparâmetro e a matriz de confusão depois da implementação dos hiperparâmetros.

Com hiperparâmetros:



Sem hiperparâmetros:

utilizamos estes hiperparâmetros pois acreditamos que estes são os melhores encontrados para a aumentar a acurácia dos dados.

C) No SVM, em um primeiro momento, foram utilizados os parâmetros

obtendo o resultado

```
parameters_svm = {#'float':range(1,50),
                 'kernel':['linear','poly','rbf','sigmoid','precomputed'],
                 'gamma':['scale','auto'],
                 'degree':range(1,50)
```

com o auxílio do _____ obtemos os melhores hiperparâmetros, que elevou nossa acurácia em x%

Sendo assim, os resultados comparados podem ser observados abaixo na matriz de confusão antes do hiperparâmetro e a matriz de confusão depois da implementação dos hiperparâmetros.

utilizamos estes hiperparâmetros pois acreditamos que estes são os melhores encontrados para a aumentar a acurácia dos dados.

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.

Em um primeiro momento, foi decidido como parâmetro de avaliação de desempenho a acurácia dos modelos de aprendizagem de máquina, visto que essa métrica engloba todas possíveis variáveis de resultado, como os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Dessa forma, com base nos resultados finais obtidos nos testes, conclui-se que o melhor método de aprendizagem de máquina para este projeto até o momento é o KNN, com uma acurácia de 47%.

Esse valor indica que , dentre todos os resultados dos testes de predições, o número de valores verdadeiros positivos e negativos representam 47% do total, o que sugere uma revisão nos tratamentos dos dados de modo a eliminar os falsos positivos e negativos, responsáveis por abaixar essa porcentagem e diminuir a precisão e recall do modelo, o que consequentemente abaixam a acurácia de modo geral.

Além disso, acredita-se que esse resultado possa ser ainda otimizado baseado em outros modelos de classificação mais sofisticados, como o svm, que deverá ser implementado futuramente visando buscar melhores resultados e métricas de comparação.

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:

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.