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

27/09/2002	Jordan Andrade	2.1	Revisão e pequenas correções nos tópicos 4.4 e 4.5
04/10/2022	Dayllan Alho	2.2	Revisão dos tópicos 4.4 e 4.5
05/10/2022	Dayllan Alho	2.3	Acréscimo de informação nos tópicos 4.4, 4.5 e 5.0
05/10/2022	Jackson Aguiar	2.4	Acréscimo de informação nos tópicos 3.1, 3.2 e 3.3
05/10/2022	Gabriel Nhoncanse	2.5	Acréscimo de informações nos tópicos 3.2 e 3.3

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

3.1. CRISP-DM

Buscando a melhor abordagem a problemática de fluxo de trabalho sobre a mineração de dados, fomos direcionados a utilizar e a compreender o método CRISP-DM, que se baseia em uma rotina de atividades com o cumprimento de etapas a fim de atingir precisão em prever futuras falhas e soluções no processo.

De modo para realizar o processo de forma plena, deve-se obedecer as 6 etapas da metodologia, sendo as três primeiras etapas com o objetivo de coleta e organização dos dados a serem analisados. Elas são o entendimento do negócio, o entendimento dos dados e a preparação dos dados. Em nosso cenário, as diversas pesquisas e reuniões de entrega nos disponibilizaram conversas com os stakeholders, foram fundamentais para a obtenção do conhecimento necessário para exploração dos dados e identificação dos elementos essenciais na solução.

Após toda análise das regras que circundam o negócio e os dados, prossegue-se para as últimas três fases, a modelagem, avaliação e implementação, que assumem o papel da criação do modelo, baseando se nas etapas anteriores, e a colocação deste modelo em prática. É aqui que todo o trabalho anterior será testado e, caso necessário, refeito.

3.2. Ferramentas

Para construção do modelo, utilizamos como linguagem, o python, que nos fornece todos recursos e bibliotecas necessárias, e algumas bibliotecas que facilitaram o processo da construção do modelo e do tratamento e visualização dos dados, como o numpy, pandas, sklearn, seaborn e matplotlib. Sendo tudo executado em um ambiente virtual na nuvem, o Google Colab, assim removendo a necessidade de hardware equivalente, além de permitir o acesso e modificação dos demais integrantes, criando um espaço para o versionamento do modelo.

Além disso, utilizamos o miro como ferramenta para os frameworks de negócio, possibilitando que pudéssemos conhecer melhor nosso público-alvo e a empresa parceira, assim como o mercado em que ela está inserida, e, com isso, desenvolvemos uma solução mais assertiva para o contexto de sua criação.

3.3. Principais técnicas empregadas

Inicialmente, excluímos colunas que tivessem menos de ⅓ dos campos de dados com registros válidos, ignorando o restante devido à ausência de registros significativos. Após isso, continuamos tratando os dados com o uso de estratégias como “label encoding” e “one hot encoding”. Além disso, construímos algoritmos de tratamento para dados, como as funções

"*convertDate*" e "*pegar_idade*", que colaboram em formatar as datas e definir a idade como saída para o dataset. Visando a obtenção de resultado mais precisos, categorizamos a "*class_sobrevida*" com um algoritmo de loop, sendo necessário transformar a coluna dos dias de sobrevida com um filtro, de 1 a 5, para que os dados pudessem ser melhor enquadrados dentro da aplicação do modelo e evitasse ruídos nos resultados. Também renomeamos todas as colunas restantes, a fim de facilitar o entendimento do significado de cada uma. Por fim, criamos uma nova tabela, contendo apenas as informações das pacientes que já haviam morrido, sendo assim, treinamos o modelo apenas com casos em que era presente um dado de sobrevida no banco.

Com os dados tratados, definimos o que seria usado para treinar o modelo e o que seria usado para teste, separando também as colunas que seriam usadas para "perguntar" ao modelo e qual seria a coluna "resposta", ou coluna "target", sendo ela a que o modelo prediria os valores. Após isso, testamos os modelos que teríamos no projeto, sendo eles: Regressão linear, árvore de decisão, knn, svm e random forest.

Tendo os modelos prontos, buscando aumentar a acurácia deles, testamos eles com o uso de hiperparâmetros, com o auxílio de Random Search e Grid Search.

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

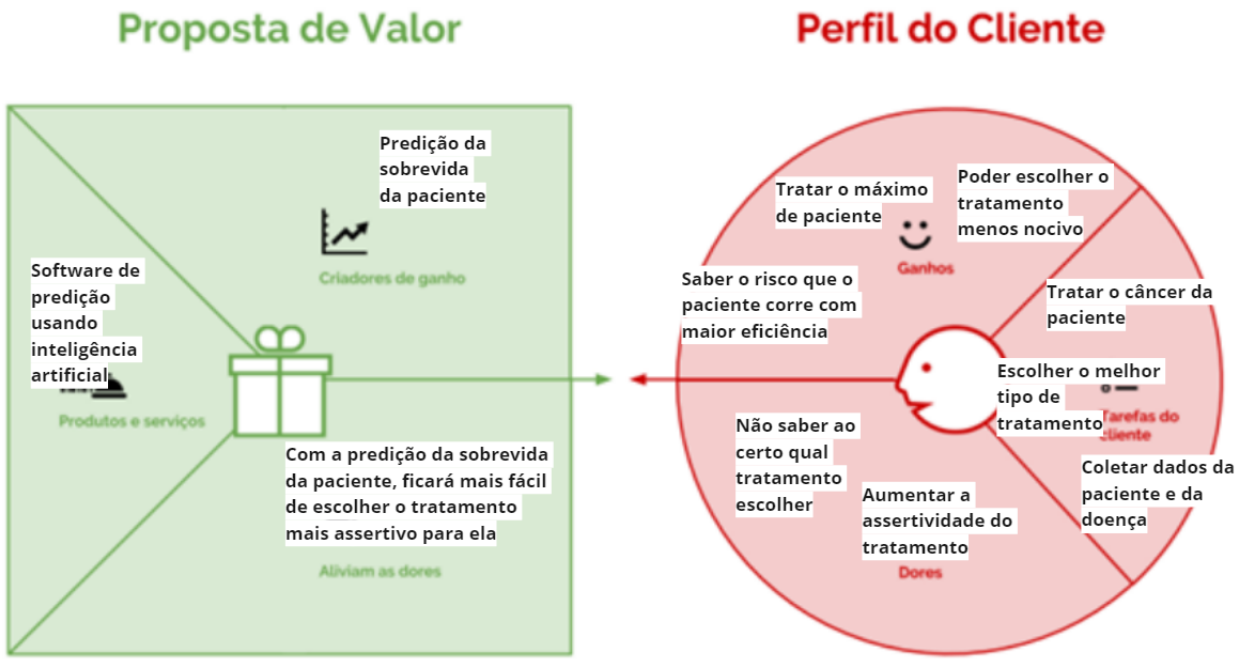


Imagem 2: Value Proposition Canvas

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:

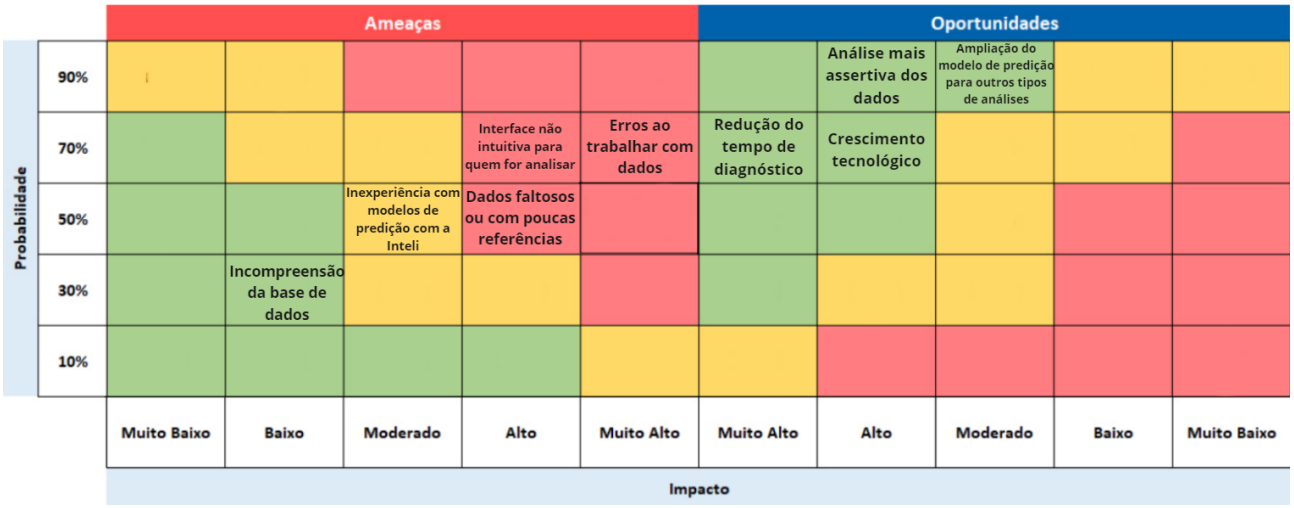


Imagem 3: Matriz de riscos

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

4.4.1. Modelos testados e comparados

Durante o desenvolvimento da Sprint 4, decidimos coletar alguns modelos de predição para melhor visualizar a dinâmica de treinamento e teste do algoritmo. Para isso, utilizamos seis grandes modelos, sendo eles: o KNN (KNowledge Now), a Árvore de Decisão, o SVM, a Regressão Linear,a Regressão Logística e o Random Forest. 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, e portanto removemos esses dois modelos da nossa documentação antes mesmo de iniciarem os teste com cada modelo.

Assim, durante a modelagem selecionamos quatro 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.

4.4.1.2. KNN (KNowledge Now)

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. o Cientista de Dados da Experian DataLabs e mestrando da UNIFESP, Willian Dihanster afirma que o KNN é:

> o algoritmo k Vizinhos Mais Próximos, conhecido como kNN (do inglês, k Nearest Neighbors). Na primeira parte do artigo, irei introduzir a teoria do algoritmo, e na segunda parte, farei uma implementação simples em Python. (DIHANSTER, Willian. 2020)

Uma demonstração que representa bastante o KNN é demonstrada a seguir:

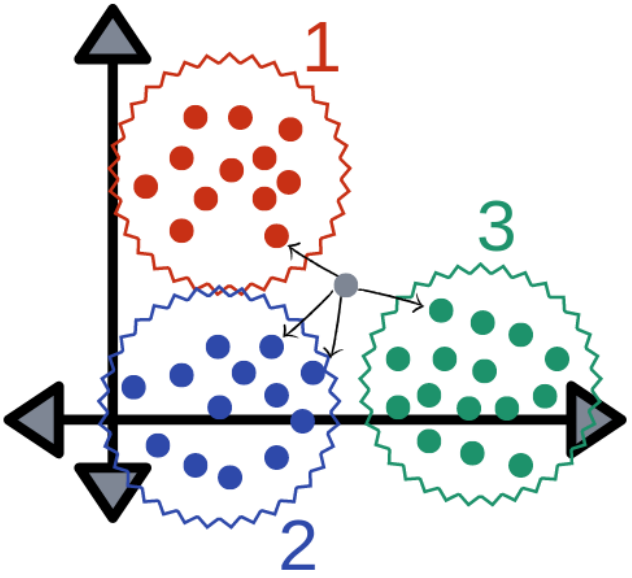


Imagem 7: exemplo de KNN

Na imagem acima é possível ver que a representação circular em tons de cinza é nosso algoritmo e que ele se baseia no conjunto de vizinhos para determinar qual quem ele é, ou seja, por comparação. o KNN funciona desta forma, observando o conjunto de vizinhos a sua volta e determinando quem eres.

Deste modo, percebemos que 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.

```
1 #testando o modelo
2 model_knn= KNeighborsClassifier(n_neighbors=23)
3 model_knn.fit( X_train, Y_train)
4
5 Y_pred_knn = model_knn.predict(X_test)
```

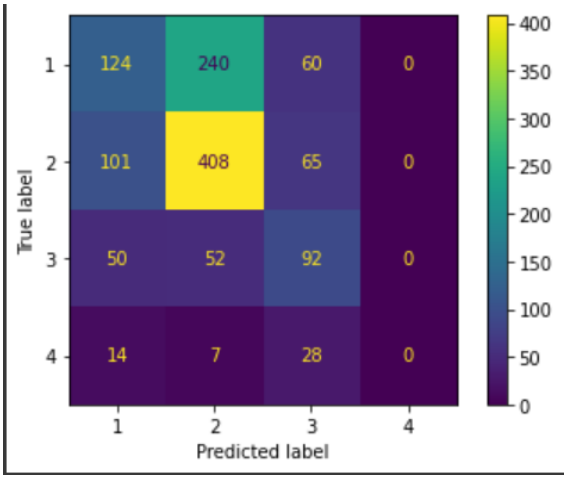
Imagem 8: testando o modelo de classificação do KNN

Nesse caso, já tínhamos definido no código as variáveis que iremos usar, sendo elas:

```
#separando os dados:
x_entrada = x_inter3[['tumor','nodulo',
                      'metastese', 'tempo_de_amamentacao',
                      'primeira_menstruacao', 'fumante', 'alcolatra',
                      'grau_parentesco_1', 'grau_parentesco_2', 'grau_parentesco_3',
                      'tipo_tumor', 'idade']].values
y_saida = x_inter3['sobrevida'].values

X_train, X_test, Y_train, Y_test = train_test_split(x_entrada, y_saida,
                                                    test_size = 0.3,
                                                    random_state = 42)
```

Imagem 9: Separando dados de treino e teste para todos os modelos

Assim, a acurácia de treino foi de 58%, e a de teste 50%. Abaixo pode ser vista a acurácia, representada pela sigla Acc, tanto de treino quanto a de teste.




```
Acc treino:  0.5831316972001382
Acc teste:  0.5028203062046737
Revocação: [0.29245283 0.71080139 0.4742268  0.        ]
Precisão: [0.42906574 0.57708628 0.3755102  0.        ]
F1_score: [0.42906574 0.57708628 0.3755102  0.        ]
```

Imagem 10: Matriz de confusão do KNN

Como pôde ser percebido, o modelo sem o uso de qualquer hiper-parâmetro possui uma acurácia de teste de 50% , indicando uma forte taxa de erro do modelo em cima de predições falso negativas e positivas, visto que a quantidade de desses tipos de erros é inversamente proporcional à pontuação da acurácia.Entretanto, ainda que possuam uma considerável taxa de erro, ainda mostra um desempenho melhor que o modelo anterior para esse mesmo tipo de teste.

4.4.1.3 Árvore de Decisão

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.

As árvores de decisão são métodos de aprendizado de máquinas supervisionado não-paramétricos, sendo muito utilizados em tarefas de classificação e regressão em machine learning. A sua representação gráfica é formada por informações denominadas como nós. Estes se ligam e direciona “perguntas” com possíveis “respostas”. para além disso, toda árvore possui uma raiz, no caso da árvore de decisão, o nó principal é a nossa raiz, este possui o maior nível hierárquico e é de onde parte os outros ramos/elementos, que são seus filhos, onde estes podem ter seus próprios filhos e assim por diante.Ao terminar o número de filhos, a última ramificação desta árvore é chamada de folha ou terminal e geralmente, é representada por um símbolo arredondado.

Compreendendo a estrutura da árvore, o principal a se destacar é que a árvore armazena as regras de cada nó, ou seja, a regra de cada pergunta e de cada resposta para a tomada de decisão.

Abaixo temos uma representação gráfica de uma árvore de decisão construída a partir do exemplo do profissional Rafael Campos, cientista de dados da empresa Hekima:

> Em uma árvore de decisão, uma decisão é tomada através do caminhamento a partir do nó raiz até o nó folha. Para elucidar melhor, suponha que tenhamos essa situação: você vai fazer uma reunião com seus amigos em sua casa, e vai rolar comidas e bebidas. Você quer jogar um jogo de tabuleiro, mas não sabe qual o melhor para essa situação, portanto, você vai recorrer a árvore de decisão para tomar sua decisão de forma mais correta. Para isso vamos caminhar por ela a partir do nó raiz, que contém a pergunta: Está jogando com crianças? Não, então iremos para nó filho da esquerda. Vamos jogar por mais de duas horas? Acho que sim! Assim, o próximo é: Regras difíceis? Vou tomar uma, logo não vou querer pensar muito rsrs. Resposta é não. Todos os jogadores vão ficar até o final? Vamos supor que sim! Desse modo, chegamos ao nó folha, com jogo Le Havre. (CAMPOS, Rafael. 2017).

A representação gráfica do jogo perfeito do exemplo do Rafael Campos está sendo representada abaixo.

Imagem 11: Exemplo de árvore de decisão.

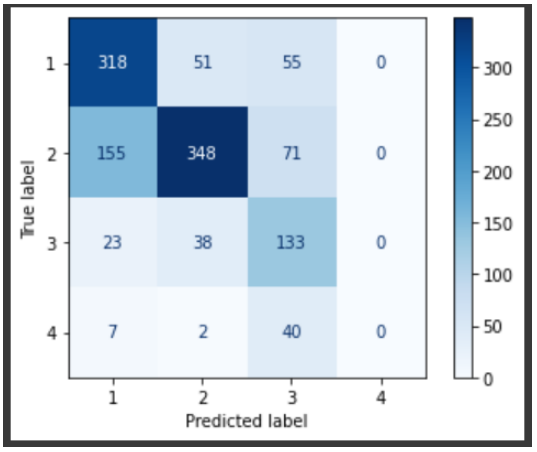
Nesta representação é possível ver representações em formatos diferentes, principalmente por demonstrar onde está o nó raiz, as ramificações da árvore e onde termina cada ramificação.

No nosso modelo preditivo, utilizamos o seguinte treinamento da árvore de decisão:

```
# # Treinando o modelo
model_tree = DecisionTreeClassifier(criterion='entropy', random_state = 42)

model_tree.fit(X_train,Y_train)

# Fazendo as predições
Y_pred_tree = model_tree.predict(X_test)

```

Imagem 12: Treinamento do modelo da árvore de decisão.

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.




```
Acc treino:  0.9913584514344971
Acc teste:  0.5431103948428686
Revocação: [0.55896226 0.6097561  0.39690722 0.20408163]
Precisão: [0.5361991  0.6294964  0.385      0.23255814]
F1_score: [0.5361991  0.6294964  0.385      0.23255814]
```

Imagem 13: Matriz de Confusão da Árvore de Decisão

Como pôde ser percebido acima, o modelo sem o uso de qualquer hiper parâmetro possui uma acurácia de teste de 54% , indicando uma forte taxa de erro do modelo em cima de predições falso negativas e positivas, visto que a quantidade de desses tipos de erros é inversamente proporcional à pontuação da acurácia.

4.4.1.4. SVM

SVM: O SVM, em tradução livre, é a máquina de vetor de suporte, são algoritmos de predição, que detectam, regridem e classificam valores discrepantes que são adaptáveis. Esse algoritmo treina os dados rotulados e gera um hiperplano que categoriza novos exemplos. Ele gera uma linha que divide um plano em duas partes, onde cada uma se situa de um lado, ou seja, o que foi treinado e o que é predito.

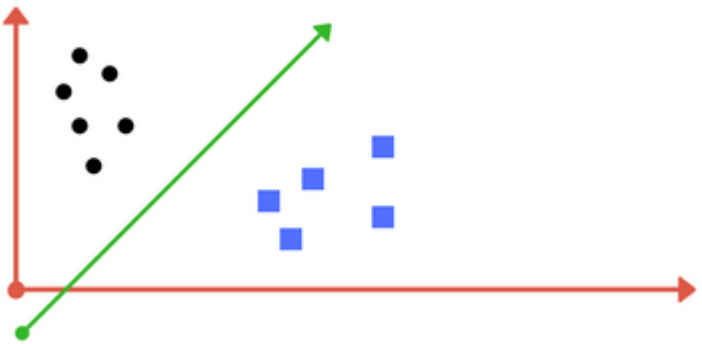


Imagem 14: Exemplo explicativo do SVM

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

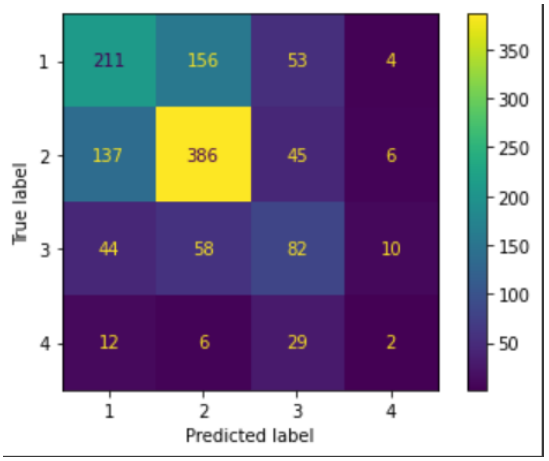
escolhe a reta entre dois grupos que se distancia mais de cada um e, após descobrir essa reta, o programa conseguirá predizer a qual classe pertence um novo dado ao checar de qual lado da reta ele está.

O código usado para treinar o modelo foi representado abaixo, juntamente com a definição dos dados:

```
#testando o modelo
x = x_inter3[['tumor','nodulo','metastese', 'tempo_de_amamentacao', 'primeira_menstruacao',
                      'fumante', 'alcolatra', 'grau_parentesco_1', 'grau_parentesco_2', 'grau_parentesco_3',
                      'tipo_tumor', 'idade']].values
y = x_inter3['sobrevida'].values

#separação do que será usado para treinar a máquina e o que será usado para testá-la
x_treino, x_teste, y_treino, y_teste = train_test_split(x,y,
                                                        test_size = 0.3,
                                                        random_state = 42)


clf = svm.SVC(kernel='poly', C = 1.0)
clf.fit(x_treino, y_treino)
y_pred = clf.predict(x_teste)
clf.score(x_teste, y_teste)
```

Imagem 15: treinando e testando modelo do SVM, juntamente com os dados utilizados.

Com as variáveis (selecionadas previamente) já inseridas no modelo, conseguimos alcançar 51% de score para a acurácia.




```
Acc treino:  0.5523677843069478
Acc teste:  0.5181305398871877
Revocação: [0.22877358 0.67421603 0.81958763 0.        ]
Precisão: [0.52150538 0.63546798 0.35650224 0.        ]
F1_score: [0.52150538 0.63546798 0.35650224 0.        ]
```

Imagem 16: Matriz de confusão do SVM

Esse resultado indica uma forte taxa de erro do modelo em cima de predições falso negativas e positivas, visto que a quantidade de desses tipos de erros é inversamente proporcional à pontuação da acurácia. Entretanto, a baixa acurácia (fortemente influenciada pelos falsos positivos) e alta revocação(fortemente influenciada pelos falsos negativos) indica que o modelo erra mais os falsos positivos que os falsos negativos, ainda que o valor exorbitantemente alto da revocação (100%) sinalize para um vício do modelo durante o treinamento.

4.4.1.5. Random Forest

Random forest: A ideia principal do Random Forest é fazer o agrupamento de uma série de árvores de decisão criadas durante o processo de treinamento do modelo gerar predições (nesse caso classificatórias) com base nas comparações conjuntas entre as estruturas arbóreas desenvolvidas.

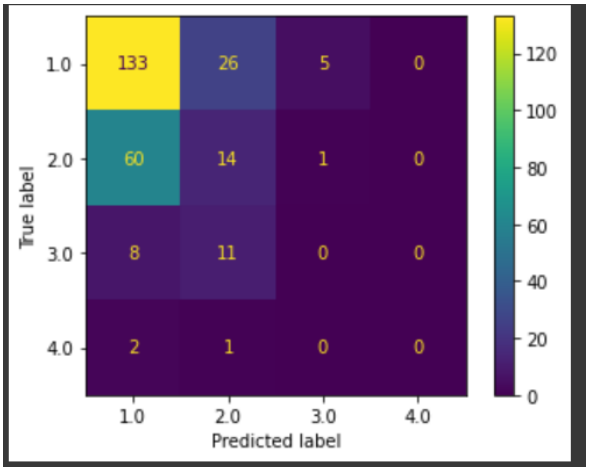
Esse modelo foi utilizado como mais uma métrica de comparação entre qual modelo classificatório seria o mais adequado para o problema em análise.

Utilizamos o código abaixo para gerar o treinamento e teste do modelo em questão

```
[ ]  1 #testando o modelo
     2 model_rf= RandomForestClassifier(random_state=3,n_estimators=100, max_depth = 30, criterion = 'gini')
     3 model_rf.fit( X_train, Y_train)
     4
     5 Y_pred_rf = model_rf.predict(X_test)
```

Imagem 17: Teste do modelo Random Forest.

Com as variáveis (selecionadas previamente) já inseridas no modelo, conseguimos alcançar 56% de score para a acurácia.




```
Acc treino:  0.9950738916256158
Acc teste:  0.5632183908045977
Revocação: [0.81097561 0.18666667 0.
Precisão: [0.65517241 0.26923077 0.
F1_score: [0.65517241 0.26923077 0.
```

Imagem 18: Matriz de Confusão do Random Forest

Esse resultado indica uma forte taxa de erro do modelo em cima de predições falso negativas e positivas, visto que a quantidade de desses tipos de erros é inversamente proporcional à pontuação da acurácia. Entretanto, a baixa acurácia (fortemente influenciada pelos falsos positivos) e alta revocação(fortemente influenciada pelos falsos negativos) indica que o modelo erra mais os falsos positivos que os falsos negativos.

4.4.2. Hiperparâmetros

Buscando obter as melhores métricas de desempenho para os modelos acima, foram utilizadas as ferramentas de busca de hiperparâmetros Grid Search e Randomized Search, de modo a obter 2 resultados distintos de uma ferramenta para a outra para que pudesse ser realizado uma comparação entre os resultados obtidos.

4.4.2.1. KNN com hiperparâmetros

4.4.2.1.1 KNN com Grid Search

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

```
parameters_knn = {'n_neighbors':range(5,110,5),
                  'weights':['uniform'],
                  'algorithm':['auto'],
                  'leaf_size':range(1,50)
```

Imagem 19: Parâmetros utilizados para o KNN

Cada um destes parâmetros possui uma explicação e características diferente, sendo elas:

- N_neighbors: É o número de amostras vizinhas utilizadas pelo modelo para a comparação;
- Weights: É o peso que cada dado da amostra possui sobre a decisão final do modelo, podendo ser ele uniforme("uniform") para todos os dados ou inversamente proporcional a distância da amostra analisada("distance").
- Algorithm: Qual o tipo de algoritmo será utilizada para a testagem do modelo, podendo ser algoritmo "ball_tree","kd_tree" ou "brute".
- Leaf_size: É o tamanho da folha passado para o parâmetro "ball_tree" ou "kd_tree".

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

```
grid_knn = GridSearchCV(KNeighborsClassifier(), parameters_knn)
# Treina os modelos e guarda na variável modelGS o melhor modelo
grid_knn.fit(X_train, Y_train)

modelGS_knn= grid_knn.best_estimator_

y_pred_knn = modelGS_knn.predict(X_test)

accuracy = accuracy_score(Y_test, y_pred_knn)
```

Imagem 20: Treinamento e teste com Grid Search

Utilizamos ambas as ferramentas de busca das melhores métricas foram treinadas e os códigos decidimos comparar os modelos e visualizar a diferença plotada delas, assim, abaixo há a matriz de confusão do modelo do KNN com o Grid Search e Randomized Search, respectivamente

4.4.2.1.2 KNN com Randomized Search

Buscando ainda analisar outros modelos de busca de hiperparâmetros para servir de comparação com os resultados obtidos, também foi utilizado o Randomized Search para a busca dos melhores hiperparâmetros.

Dessa forma, decidimos testar diferentes valores e características para os mesmos parâmetros mencionados acima, utilizando o código a seguir:

```
parametros_knn={
    'n_neighbors':range(5,110,5),
    'weights':['uniform','distance'],
    'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'],
    'leaf_size':range(5,50,5),
}
randomized_search_knn = RandomizedSearchCV(estimator = KNeighborsClassifier(),param_distributions=parametros_knn)
randomized_search_knn.fit(X_train,Y_train.squeeze())

modelRS_knn = randomized_search_knn.best_estimator_

y_predRS_knn = modelRS_knn.predict(X_test)

accuracy = accuracy_score(Y_test, y_predRS_knn)
```

Imagem 21: Buscando as melhores métricas com o Randomized Search.

A partir da geração destes dois modelos conseguimos visualizar a acurácia e a matriz de confusão destes modelos

Com Grid Search

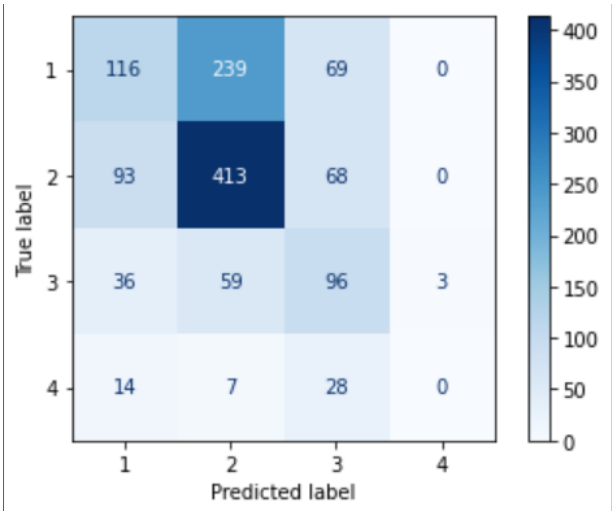


Imagem 22: Matriz de Confusão com o Grid Search

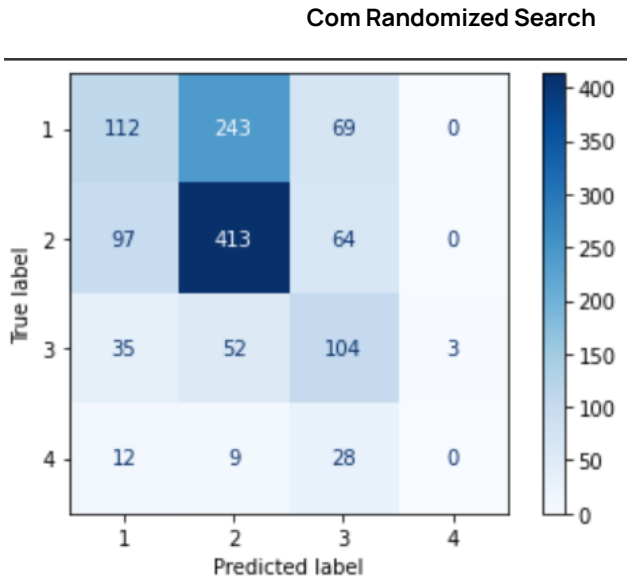
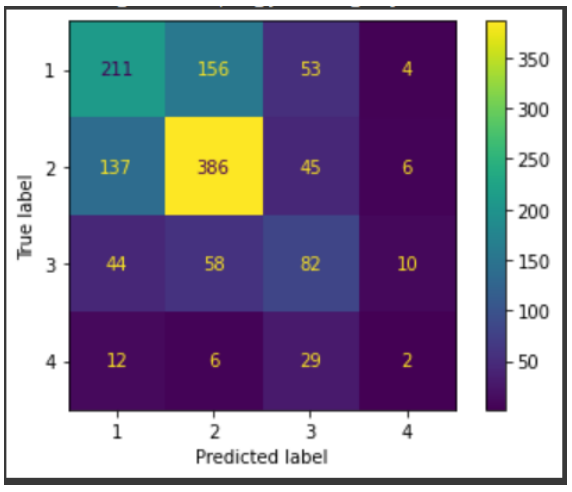


Imagem 23: Matriz de Confusão com o Randomized Search

Esses resultados mostram mudanças pouco significativas das métricas em comparação ao modelo sem hiper-parâmetros, sugerindo uma enorme presença de ruídos durante a limpeza dos dados que dificulta a aprendizagem do modelo mesmo com hiperparâmetros.

4.4.2.2. SVM com hiperparâmetros

No SVM, em um primeiro momento, obtivemos o seguinte resultado: (Sem hiperparâmetros).




```
Acc treino:  0.5523677843069478
Acc teste:  0.5181305398871877
Revocação: [0.22877358 0.67421603 0.81958763 0.        ]
Precisão: [0.52150538 0.63546798 0.35650224 0.        ]
F1_score: [0.52150538 0.63546798 0.35650224 0.        ]
```

Imagem 24: Matriz de confusão do SVM sem Hiperparâmetros.

Com o auxílio do Grid Search, realizamos diversos testes com base nos hiperparâmetros pré-selecionados e retornamos o melhor resultado dentre todas essas repetições do modelo. Utilizamos os seguintes hiperparâmetros:

- Kernel: É um tipo de método para classificação linear para resolver um problema não linear. Pode ser do tipo “rbf”, “poly”, “sigmoid” e “precomputed”.
- Gama: É o coeficiente de kernel para os tipos “rbf”, “poly” e “sigmoid”. Pode receber os valores “scale” ou “auto”
- Degree: É o grau da função kernel polinomial.

Sendo assim é possível visualizar os métodos plotados abaixo:

```
parameters_svm = {
                 'kernel':['rbf'],
                 'gamma':['scale','auto'],
                 'degree':range(1,10)
                 }
```

Imagem 25: Métricas utilizadas

Com isso, obtivemos o resultado de 53,8% de score no modelo, comparado ao de 51,8% sem hiperparâmetros, sendo assim, tivemos um pequeno aumento de 2% no score do modelo.

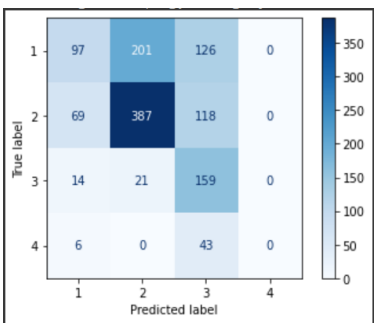
Com o auxílio do Random Search, usamos os hiperparâmetros, abaixo:

```
parametros_svm = {
    'kernel':['rbf'],
    'gamma':['scale','auto'],
    #'float':range(1,50),
    'degree':range(1,5)
}
```

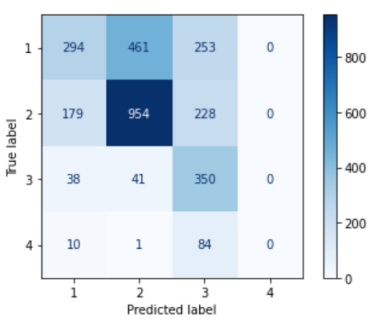
Imagem 26: Parâmetros do SVM

E com este resultado, pudemos manter o score de 53,% do modelo. 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 Grid Search:



Com Random Search:



4.4.2.3. Random Forest com hiperparâmetros

No Random Forest, em um primeiro momento, foram utilizados os seguintes parâmetros:

```
model_rf= RandomForestClassifier(random_state=3,
                                 n_estimators=100,
                                 max_depth = 30,
                                 criterion = 'gini')
```

Imagem 27: Random Forest sem definição de hiperparâmetros

Obtendo o seguinte resultado: (Sem hiperparâmetros)

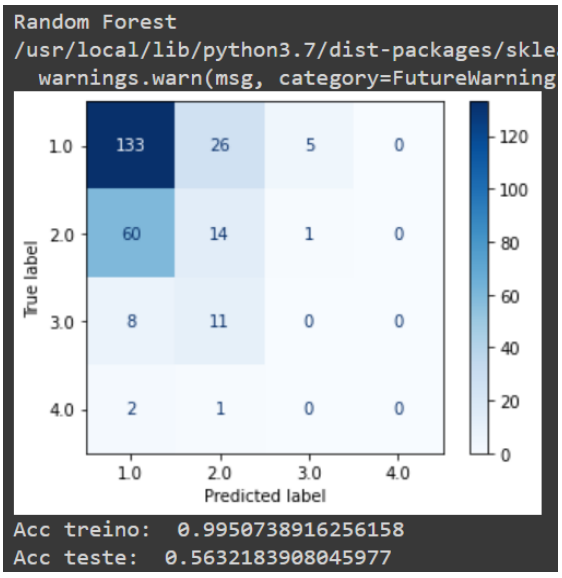


Imagem 28: Matriz de Confusão do Random Forest

E com o auxílio do Grid Search, buscamos os melhores hiperparâmetros para o nosso modelo, sendo assim, preferimos utilizá-los como método de obter o melhor score, e conseguimos um resultado de 56,3% de score no modelo. Abaixo há os hiperparâmetros que utilizamos.

```
parameters_rf = {'n_estimators':range(50,250, 50),
                 'criterion':['gini','entropy','log_loss'],
                 'min_samples_split':range(2,8,2),
                 }
```

Imagem 29: Hiperparâmetros com o Grid Search do Random Forest

Outro modelo que utilizamos para buscar os melhores hiperparâmetro foi o Random Search, que buscou o melhor para o nosso projeto e modelo, com este método, obtivemos um

resultado de 54,3% no score, o que é baixo, mas já é uma melhoria em relação ao Grid Search. abaixo é possível conferir os hiperparâmetros utilizados:

```
parameters_rf = {'n_estimators':range(50,250, 50),
                 'criterion':['gini','entropy','log_loss'],
                 'min_samples_split':range(2,8,2),
                 }
```

Imagem 30: Parâmetros utilizando o Randomized Search

4.4.2.4 Árvore de Decisão com hiper-parâmetros

Visando buscar melhores resultados para a Árvore de Decisão, foi utilizado a ferramenta de busca de hiperparâmetros Grid Search, que realiza vários testes com base nos hiperparâmetros pré-selecionados e retorna o melhor resultado dentre todas essas repetições do modelo.

Dessa forma, decidimos usar testar diferentes valores e características para os seguintes parâmetros:

- Criterion: função utilizada para a qualidade de uma divisão. Os critérios suportados são “gini” para a impureza Gini e “log_loss” e “entropia” ambos para o ganho de informação de Shannon, que diz que a quantidade de informação de um evento A depende apenas da probabilidade p(A) desse evento, e é tanto maior quanto menor for a probabilidade.

- Splitter:estratégia utilizada para escolher a divisão em cada nó. As estratégias suportadas são “best” para escolher a melhor divisão e “random” para escolher a melhor divisão aleatória.

- Max_depth:representa a profundidade máxima da árvore.Se Nenhum, os nós são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos que o número mínimo de amostras.

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

Imagem 31: parâmetros utilizados para a Árvore de Decisão

```
Acc treino: 0.6657449014863464
Acc teste: 0.6438356164383562
Revocação: [0.75       0.60627178 0.68556701 0.        ]
Precisão: [0.63220676 0.79271071 0.44481605 0.        ]
F1_score: [0.68608414 0.68706811 0.53955375 0.        ]
```

Imagem 32: Resultado obtido com os hiperparâmetros da Árvore de Decisão

Como pôde ser visualizado, utilizando dos melhores hiper-parâmetros para o modelo segundo o Grid Search, a acurácia teve um aumento de 7% em relação ao modelo anterior e, visto o aumento da pontuação do recall, que é fortemente influenciado pela quantidade de falsos negativos, acredita-se que esse tipo de dado tenha diminuído de um modelo para outro.

4.5. Avaliação

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 a Árvore de Decisão, com uma acurácia de aproximadamente 64%, utilizando as métricas do Grid Search. Esse valor indica que, dentre todos os resultados dos testes de predições, o número de valores verdadeiros positivos e negativos para os modelos representam cerca de 64% do total, o que sugere uma revisão no 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.

Esse valor indica que , dentre todos os resultados dos testes de predições, o número de valores verdadeiros positivos e negativos para os modelos representam 64% do total, o que sugere uma revisão no 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.

Ademais, vale ressaltar que foram desconsiderados os resultados de acurácia dos modelos cujo resultado foi igual a 100% sem que esse valor fosse compatível com as outras métricas, indicando o possível vício de treinamento e fortes ruídos em uma ou mais features durante a limpeza e manipulação dos dados.

Em suma, recomenda-se que seja realizada uma revisão no processo de feature engineering para buscar otimizar ainda mais os resultados para assertividade do modelo, além

de avaliar os pesos que cada coluna possui sobre a classificação de cada modelo de treinamento de modo a buscar e eliminar possíveis dados viciantes.

5. Conclusões e Recomendações

Acreditamos que o tratamento e escolha das features pode ser feito de diversas maneiras, e destacamos que selecionamos uma estratégia de escolha e que o nosso modelo caminha nesta direção de predição. É viável reestruturar a base de dados e rodar novamente as linhas de código para fazer o mesmo tipo de tratamento, apenas mudando o caminho da tabela tratada, isso possivelmente gerará uma melhor acurácia dos modelos, desde que a solução da predição seja a predição do tempo de sobrevida, isso porque em todos os modelos testados os dados plotados já estão encaminhados para o tempo de sobrevida, o filtro de dias em 1, 2, 3 e 4 e o filtro de cores seguem essa regra de negócio.

Juntamente com nosso código, salientamos que este modelo não é uma solução de substituição da mão-de-obra de qualquer funcionário desta ou de outra instituição. Ressaltamos que nosso modelo preditivo tem o foco em aprimorar o trabalho destas pessoas e não substituí-las, tal visto que, além de ser um projeto de Machine Learning, este projeto é um “humanizador de máquinas”, ou seja, não são apenas números que o modelo prediz, mas ele prediz a esperança de dias melhores para cada paciente que é o foco do nosso projeto. Por fim, este modelo pode e deve ser utilizado e ampliado de forma abrangente para predizer o tempo de sobrevida do paciente com câncer de mama.

6. Referências

INCA. *Câncer de mama.* s.d. https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/mama (acesso em Setembro de 2022).

—. *Câncer de mama: saiba como reconhecer os 5 sinais de alerta.* s.d. https://www.gov.br/saude/pt-br/assuntos/saude-brasil/eu-quero-me-exercitar/noticias/2021/cancer-de-mama-saiba-como-reconhecer-os-5-sinais-de-alerta (acesso em 2022 de Agosto).

Python. *Python Software Foundation.* s.d. https://www.python.org/.

—. *The Python Standard Library.* s.d. https://docs.python.org/3/library/ (acesso em Outubro de 2022).

https://medium.com/dftblog/knn-introdu%C3%A7%C3%A3o-aos-algoritmos-de-aprendizado-de-m%C3%A1quina-dd2107693651 ‹acessado em: 04/10/2022.

https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69 ‹acessado em: 04/10/2022.