UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Pesquisa sobre as técnicas de Inteligência Artificial Explicável: Uma perspectiva do usuário

Fábio Luiz Daudt Morais

Orientador

Ana Cristina Bicharra Garcia

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 2020

Pesquisa sobre as técnicas de Inteligência Artificial Explicável: Uma perspectiva do usuário

FÁBIO LUIZ DAUDT MORAIS

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Dr. ANA CRISTINA BICHARRA GARCIA — UNIRIO

Dr. SEAN WOLFGAND MATSUI SIQUEIRA — UNIRIO

Dr. RODRIGO CLEMENTE THOM DE SOUZA — UFPR

Dr. DANIEL LOPES CINALLI — TECHNIPFMC

Dr. PAULO SÉRGIO MEDEIROS DOS SANTOS — UNIRIO

RIO DE JANEIRO, RJ - BRASIL
JULHO DE 2020.

Daudt Morais, Fábio Luiz.
B118 Pesquisa sobre as técnicas de Inteligência Artificial explicável: Uma perspectiva do usuário / Fábio Luiz Daudt Morais, 2020
xiii, 176f.

Orientador: Ana Cristina Bicharra Garcia 1.6cmDissertação (Mestrado em Informática) - Universidade Federal do
Estado do Rio de Janeiro, Rio de Janeiro, 2020.

1. Machine Learning. 2. Explainable Artificial Intelligence. 3. Interpretable Machine Learning.
4. Artificial Intelligence. 5. Human-grounded evaluation.
I. Bicharra Garcia, Ana Cristina. II. Universidade Federal do Estado do
Rio de Janeiro (2017-). Centro de Ciências Exatas e Tecnologia. Curso de
Mestrado em Informática. Título.

CDD - 004.678

Agradecimentos

Agradeço primeiramente a Deus por me dar a oportunidade de concluir o mestrado e assim me permitir alcançar mais um degrau em minha vida acadêmica e profissional.

Meus sinceros agradecimentos a minha filha, Nina; a minha mãe, Cleuza; aos meus irmãos, Caio e Bárbara e a minha namorada, Carolina, por todo apoio, carinho e por estarem ao meu lado nessa longa jornada. Meus agradecimentos aos colegas de trabalho da Fiocruz por toda compreensão, ajuda e amizade que muito me ajudou a alcançar mais essa etapa em minha vida.

Agradecimento especial à minha orientadora Ana Cristina Bicharra Garcia, por toda paciência, tutoria e ensinamento. Sei que não fui aquele aluno excepcional, então minha eterna gratidão por todo auxílio e tempo doado nessa caminhada em busca do conhecimento.

Meus agradecimentos ao Paulo que me ajudou no método qualitativo e ao Daniel Cinalli que me auxiliou em todo o processo de pesquisa. Sou muito grato à Deus por ter tido a oportunidade de ter cruzado com pessoas tão generosas como vocês. A mentoria, ajuda, incentivo e suporte que vocês me deram foi incondicional. Obrigado, obrigado e obrigado!!!

Agradeço também aos professores que eu tive a oportunidade de conhecer e ser aluno nessa jornada. E a Unirio, corpo docente e administrativo, por oferecer ensino público com excelência.

Aos colegas de jornada, em especial (Carolina Sacramento, Wagner Silva, Miriam Oliveira e Jomar Silva), que compartilharam suas experiências, me incentivaram e ajudaram a vencer todos os obstáculos que apareceram no meu caminho ao longo desses anos. Muito obrigado!!!

"Infalível Criador, que, dos tesouros da Vossa sabedoria, tirastes as hierarquias dos

anjos, colocando-as com ordem admirável no céu; Vós, que distribuístes o universo com encantadora harmonia; Vós, que sois a verdadeira fonte da luz e o princípio supremo da sabedoria, difundi sobre as trevas da minha mente o raio do esplendor, removendo as duplas trevas nas quais nasci: o pecado e a ignorância.

Vós, que tornastes fecunda a língua das crianças, tornai erudita a minha língua e espalhai sobre os meus lábios a vossa bênção.

Concedei-me a agudeza de entender, a capacidade de reter, a sutileza de relevar, a facilidade de aprender, a graça abundante de falar e de escrever.

Ensinai-me a começar, regei-me no continuar e no perseverar até o término.

Vós, que sois verdadeiro Deus e verdadeiro homem, que viveis e reinais pelos séculos dos séculos. Amém."

(Santo Tomás de Aquino)

MORAIS, FÁBIO LUIZ E DAUDT **Pesquisa sobre as técnicas de Inteligência Artificial Explicável: Uma perspectiva do usuário**. UNIRIO, 2020. XX páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

A inteligência artificial (IA) se tornou parte de nossas vidas diárias, sendo comumente usadas em uma ampla gama de setores. A máquina aprende e depreende de dados e da interação conosco e com o ambiente. Os esforços de pesquisa têm sido em produzir sistemas precisos, confiáveis que gerem resultados robusto. No entanto, a maioria de nós tem pouco ou mesmo nenhum entendimento e conhecimento sobre como esses sistemas inteligentes funcionam e geram resultados e ações que podem impactar nossas vidas. A maioria dos algoritmos usados para o aprendizado de máquina não podem ser examinados e interpretados por humanos, funcionando como caixas pretas, levando a um comportamento de "pegar ou largar o resultado". A aceitação e inserção da tecnologia em nossas vidas também leva a questionamentos de não se aceitar sem entender. A área de inteligência artificial explicável (XAI - Explainable Artificial Intelligence) é nova e tem por objetivo trabalhar essas questões relacionadas à abertura da caixa preta do raciocínio no agente inteligente, seja para aumentar a confiança nos resultados, para auditar o processo prevenindo vieses e preconceitos, para responsabilizar pelas consequências dos resultados e mesmo para deixar transparente o processo de raciocínio. Tudo isso passa pelo entendimento e aceitação do processo computacional pelos clientes e donos desses sistemas. Este trabalho avalia a compreensibilidade das explicações de diferentes técnicas de XAI, na perspectiva dos especialistas do domínio, com o objetivo de medir o entendimento dos resultados, a confiança no sistema e a aceitação da explicação produzida. Para tanto, desenvolvemos um sistema diagnóstico de câncer, com alta acurácia, usando métodos ensembles a partir de uma base de dados pública. Rodamos alguns casos e aplicamos 3 técnicas de XAI para gerar explicações. Esses casos e suas explicações foram apresentados a 12 oncologistas. Entrevistas semiestruturadas foram utilizadas para avaliar a compreensibilidade das explicações geradas pelas 3 técnicas XAI. Usamos método de pesquisa qualitativo de Teoria Fundamentada em Dados. Encontramos evidências impor-

tantes de que as técnicas atuais de XAI são informativas, mas não explicativas. Nossas observações geraram um conjunto de diretrizes e sugestões para guiar o desenvolvimento de sistemas XAI.

Palavras-chave: Inteligência artificial, Inteligência artificial explicável, Aprendizagem de máquina interpretável, Avaliação baseada no humano, Iteração humano-computador, XAI.

ABSTRACT

Artificial intelligence (AI) has become part of our daily lives, being commonly used in a wide range of industries. The machine learns and understands data and interaction with us and the environment. Research efforts have been on producing accurate, reliable systems that generate robust results. However, most of us have little or no understanding and knowledge about how these smart systems work and generate results and actions that can impact our lives. Most of the algorithms used for machine learning cannot be examined and interpreted by humans, functioning as black boxes, leading to "take or leave the result"behavior. The acceptance and insertion of technology in our lives also leads to questions about not accepting yourself without understanding. The Explainable Artificial Intelligence (XAI) area is new and aims to address these issues related to opening the black box of reasoning in the intelligent agent, either to increase confidence in the results, to audit the process, preventing bias and prejudice , to be responsible for the consequences of the results and even to make the reasoning process transparent. All this goes through the understanding and acceptance of the computational process by the customers and owners of these systems. This work assesses the comprehensibility of the explanations of different XAI techniques, from the perspective of experts in the field, with the aim of measuring the understanding of the results, the confidence in the system and the acceptance of the explanation produced. To this end, we developed a cancer diagnosis system, with high accuracy, using ensemble methods from a public database. We ran some cases and applied 3 XAI techniques to generate explanations. These cases and their explanations were presented to 12 oncologists. Semi-structured interviews were used to assess the comprehensibility of the explanations generated by the 3 XAI techniques. We use qualitative research method of Grounded Theory. We found important evidence that current XAI techniques are informative, but not explanatory. Our observations generated a set of guidelines and suggestions to guide the development of XAI systems.

Keywords: Artificial intelligence, Explainable artificial intelligence, Interpretable machine learning, Human-grounded evaluation, Human computer interaction.

Sumário

Lista de ilustrações

Lista de tabelas

Lista de Nomenclaturas

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
DNN	Deep Neural Networks
GT	Grounded theory
IA	Inteligência Artificial
RNN	Recurrent Neural Networks
TFD	Teoria Fundamentada em Dados
XAI	eXplainable Artificial Intelligence

Lista de Símbolos e Unidades

SIGLA	UN.	SIGNIFICADO
A	(colocar a unidade)	Colocar o significado

1. Introdução

A inteligência artificial (AI, na sigla em inglês) passou por um período sombrio entre meados da década de 1970 e o início da década de 1980. Esse período ficou conhecido como inverno da inteligência artificial, pois embora a arquitetura dos sistemas inteligentes tivesse evoluído, ainda havia desafios que impediam tais seu uso em problemas reais, tais como a complexidade de processamento que tornava impraticável a espera por bons resultados e a falta de dados rotulados que guiassem o aprendizado da máquina. Esses fatores retardaram a adoção e, consequentemente, houve poucos avanços, cortes de investimentos e pouca atenção da indústria nesse período [8].

A inteligência artificial precisava se reinventar. Com o advento do big data a partir dos anos 2000 e o aumento massivo do poder computacional, as técnicas de aprendizado de máquina voltaram a ter relevância. O volume de dados gerado em variedade e velocidade cada vez maiores, permite criar modelos e atingir altos níveis de precisão. Em torno de tantas informações disponíveis e armazenadas, pode se cumprir um dos pilares da aprendizagem de máquina que é justamente a análise de dados com o objetivo de detectar padrões.

Na última década, o campo da inteligência artificial tomou força, em especial a área aprendizagem de máquina (em inglês, ML-Machine Learning) e aprendizado profundo (do inglês, Deep Learning), passou por algumas mudanças importantes. O alto volume de dados disponível e o aumento do poder computacional permitiu a evolução na qual nos encontramos hoje. Essa evolução da IA revolucionou várias áreas da computação. Por exemplo, a área de visão computacional foi totalmente impactada pelos ótimos resultados de Deep Learning no reconhecimento e mesmo geração de imagens.

Começando como um domínio puramente acadêmico e voltado para a pesquisa, vimos a adoção abrangente da inteligência artificial em uma ampla gama de setores, como saúde, finanças, educação, construção, direito e manufatura.

Atualmente, os tipos de decisões e previsões tomadas por sistemas habilitados por técnicas de IA estão se tornando muito mais profundos e, em muitos casos, críticos para a vida e o bem-estar pessoal. O que significa que em alguns domínios teremos que explicar e garantir a explicação dos resultados para aqueles que aplicarão nossos sistemas de aprendizagem de máquina.

Em alguns sistemas de tomada de decisão de alto volume, como um sistema de recomendação de varejo online e de exibição de anúncios, um algoritmo preciso é a abordagem ideal. Na maioria dos sistemas inteligentes do mundo corporativo atual, o "por que"não importa, desde que o sistema inteligente seja preciso, performático e funcione conforme o esperado.

Porém, o uso de sistemas com aprendizagem de máquina em setores críticos, como decisões de diagnóstico em sistemas de saúde pode encobrir preconceitos, vieses e mesmo imprecisões que são danosas à sociedade. Temos que ter em mente que o uso de aprendizagem de máquina implica no sistema computacional depreender relações e aprender com os dados apresentados a ele. Portanto é fundamental saber como o sistema chega aos resultados. Não há como saber se o resultado tem erros, algum potencial de injustiça ou se a decisão é razoável. Nesses cenários, o "por que"é o que mais importa.

Quando estamos avaliando um sistema, um sistema com IA também se enquadra, avaliamos a acurácia, precisão, sensitividade, especificidade e muitas outras métricas objetivas mirando resultados conhecidos. Porém tais métricas não são aplicáveis quando a questão envolve o entendimento humano dos resultados, sugestões e decisões produzidas por sistemas com aprendizagem de máquina. A falta de explicabilidade desses sistemas inteligentes impacta sua adoção e confiança neles [3]. É importante que os sistemas funcionem conforme o esperado e produzam explicações que deem transparência às razões que norteiam os sistemas inteligentes.

No contexto dos sistemas de aprendizagem de máquina, interpretabilidade ou explicabilidade significa gerar decisões nas quais um ser humano pode entender as decisões no contexto especificado, ou seja, entender a causa de uma decisão [9]. Outra definição diz que interpretabilidade é o grau em que um ser humano pode prever consistentemente o resultado de um modelo [10].

1.1 Interpretabilidade, Explicabilidade, Transparência, Compreensibilidade, Responsabilidade, Equidade e Confiabilidade

Equidade: o aprendizado não deve permitir a geração de modelos que apresentem tendências que reflitam preconceitos ou que prejudiquem algum grupo especifico de pessoas. A equidade do modelo se refere a possibilidade de consultar o modelo e descobrir interações latentes entre as características que descrevem o domínio (camada de entrada) para ter uma ideia de quais características são relevantes para gerar resultados de acordo com o modelo usado [11, 12].

Confiança: complacência em aceitar uma sugestão/resultado do sistema. Também devemos ser capazes de validar e justificar porque certas características principais foram responsáveis por conduzir determinadas decisões indicadas por um modelo [3, 13].

Responsabilidade: aferição de responsabilidade (legal e moral) em caso de consequências indesejáveis de ações ou decisões de sistemas inteligentes [14, 15].

Transparência: refere-se a possibilitar um raio X no design, no processamento do algoritmo e nos dados usados no aprendizado. A transparência algorítmica levanta questões relativas à privacidade dos dados usados para o treinamento doa algoritmo de aprendizagem de máquina [16].

Compreensibilidade: quando relacionada ao aprendizagem de máquina, refere-se à capacidade de um algoritmo de aprendizagem de representar seu conhecimento aprendido de uma maneira compreensível ao humano [7].

Interpretabilidade: revela os mecanismos internos dos algoritmos do sistema que geram os resultados. Pode ser entendida como explicação sintática do processo [17]. Por exemplo, uma explicação sintática do carro ter buzinado seria apresentar os mecanismos que levar o som do carro ter disparado revelando o mecanismo de liberação do som.

Explicabilidade: revela o modelo causal que explica o comportamento do sistema inteligente [17]. Aqui temos uma explicação semântica. Para isso é necessário se ter um modelo do que é o conhecimento. Por exemplo, uma explicação semântica para o carro ter buzinado incluiria elementos do tipo "o carro da frente estar lentamente mudando de pista sem ter ligado o pisca alerta indicando que o motorista do outro carro pode ter dormido ou estar desatento ao volante."

O termo "explicação"é usado para explicações de previsões individuais, ou seja, é como explicamos explicitamente as decisões às pessoas [9]. Interpretabilidade e explica-

bilidade são frequentemente usados como termos sinônimos na literatura de inteligência artificial explicável e são usados para descrever métodos que fornecem informações sobre o comportamento dos modelos de aprendizagem de máquina [9, 10, 18].

Em nossa pesquisa os conceitos de explicabilidade e interpretabilidade são usados como sinônimos.

A explicação deve ser inteligível para os diferentes usuários dos sistemas inteligentes que terão necessidades diferentes. A tabela 1 apresenta os diferentes consumidores potenciais de explicações dos sistemas inteligentes [7]. O foco de nossa pesquisa é o entendimento pelos especialistas do domínio. Essa escolha deve-se por serem esses os principais avalistas do sistema além de serem aqueles que podem realmente auditar o raciocínio apresentado. Consideramos que as outras perspectivas também são importantes e serão realizadas em trabalhos futuros.

Tabela 1 – Os diferentes consumidores de explicação de sistemas inteligentes e suas razões (adaptado de [7])

Audiência do XAI	Necessidade da explicação
Especialistas do domínio	Confiar no modelo e ganhar conhecimento científico
Pessoas afetadas pelo sistema	Entender sua situação perante o comportamento do sistema e verificar se os resultados são justos
Gerentes e donos da empresa	Verificar conformidade com legislação e entender as aplicações de AI na empresa
Agências Reguladoras	Certificar a conformidade do sistema com normas e legislação
Especialistas de AI, Cientista de dados e Programadores	Garantir e melhorar a eficiência do sistema e ver novas funcionalidades

1.2 Inteligência Artificial Explicável ou a necessidade de uma explicação

Um modelo de aprendizado de máquina, por si só, consiste em um algoritmo que aprende padrões e relacionamentos latentes a partir de dados sem regras fixas codificadas. Portanto, explicar como um modelo funciona para os tomadores de decisão sempre apresenta seu próprio conjunto de desafios.

A necessidade de explicabilidade surge em sistemas em que não é suficiente obter somente a previsão (o quê) [19]. Em sistemas de aprendizagem de máquina com repercussões sociais, éticos, legais, econômicas, tecnológicas, além de sistemas críticos à vida, é fundamental explique como chegou à previsão (o porquê). Podemos exemplificar essa necessidade na recente regulamentação europeia sobre o uso de dados e o direito à

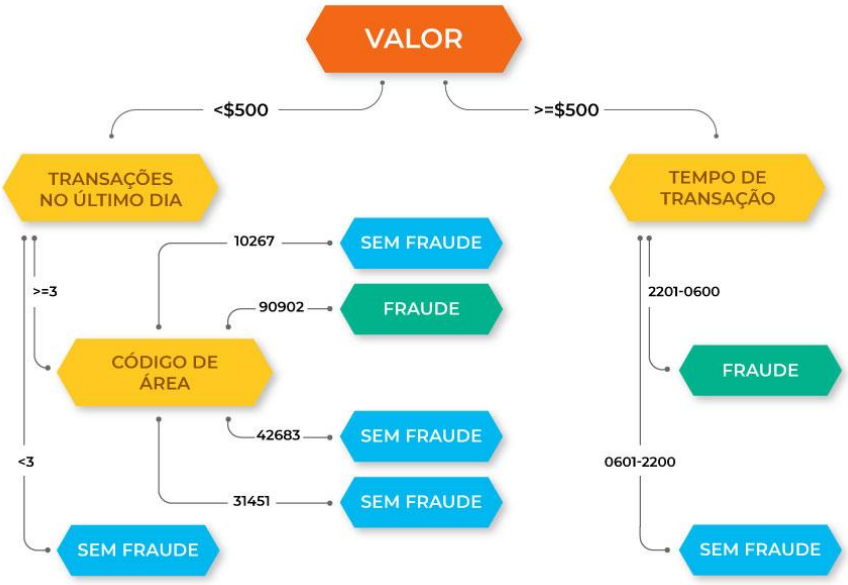

Figura 1 – Árvore de decisão. Exemplo de análise de fraude.

explicação [20] e em sistemas de diagnóstico médico.

Alguns modelos, no entanto, chamados de modelo interpretável ou modelos de caixa branca são altamente passíveis de explicação, pois todo o processo de decisão pode ser ilustrado em diagrama. Incluem modelos lineares tradicionais, baseado em árvores de decisão e sistemas de regras de negócios.

Na figura 1 temos um modelo de análise de fraude baseado em árvore em decisão. Esse modelo tem a seguinte lógica: se o valor da transação for maior que $500 e o tempo da transação estiver entre 2201-0600 significa uma fraude e se o tempo de transação estiver entre 0601-2200 não é uma fraude. Outra decisão é quando o valor da transação for menor que $500 e o número de transações nos últimos dias por menor que 3, não significa uma fraude. E se o valor da transação for menor que $500 e o código de área for o 31451 ou 42683 ou o 10267 não significa fraude, mas se o código de área for o 90902 significa uma fraude. Podemos perceber que esse modelo é altamente interpretável.

As técnicas mais complexas, como modelos ensemble, e a família de modelos de aprendizado profundo mais recente geralmente produzem melhor desempenho (já que os relacionamentos verdadeiros raramente têm natureza linear) do que as técnicas ditas interpretáveis. Mas foram construídos como caixas-pretas, ou seja, sistemas que escondem suas lógica interna para o usuário, oferecendo pouca ou nenhuma percepção discernível de como eles alcançam seus resultados. Portanto, é extremamente difícil explicar como o modelo pode realmente estar tomando suas decisões [21].

1.3 Motivação da pesquisa

Atualmente, estamos diante da barreira da explicabilidade, um problema inerente às mais recentes técnicas de aprendizagem de máquina, principalmente nos métodos ensembles, máquina de vetores de suporte e nas redes neurais profundas, que não estavam presentes na último "onda"da IA, a saber, sistemas especialistas e modelos baseados em regras.

Com o objetivo de explicar os resultados dos algoritmos de aprendizagem de máquina, então chamados de “caixa-preta”, pois até mesmo projetistas não conseguiam explicar por que o modelo chegou a uma decisão específica, além de resultados limitados pela incapacidade das máquinas de explicar suas decisões e conclusões para os seres humanos, surgiu um subcampo da inteligência artificial nomeado de inteligência artificial explicável (do inglês, explainable artificial intelligence), também conhecida pelo acrônimo XAI (eXplainable artificial intelligence) [22]. Com o avanço das pesquisas em XAI, várias técnicas, algoritmos e abordagens surgiram no decorrer dos últimos anos.

Compreensão humana, confiança e transparência são motivações populares para as técnicas de inteligência artificial explicável [23]. No entanto, temos pouco conhecimento se essas técnicas oferecem explicabilidade a seus usuários. Nesse contexto, a caracterização das explicações, ou seja, a avaliação da compreensão das explicações sob a perspectiva dos diversos tipos de usuário se faz necessária para a confiança e entendimento dos resultados providos pelas técnicas inteligência artificial explicável.

1.4 Objetivo da Pesquisa

Embora várias técnicas de inteligência artificial explicável tenham surgido nos últimos anos, não sabemos se essas técnicas entregam explicabilidade aos usuários. Ou seja, se a explicação provida por essas técnicas oferece compreensão e confiança aos diferentes usuários do sistema. Esta pesquisa foca a explicação na perspectiva dos especialistas do domínio.

Esta é uma pesquisa exploratória que visa entender se as técnicas correntes de IA explicável estão fomentando o entendimento do sistema pelos especialistas. Além de fazer uma avaliação e comparação de técnicas atuais, buscamos identificar o que falta em tais técnicas que possam guiar futuros desenvolvimentos de sistemas de IA explicável.

1.4.1 Objetivo Principal

O principal objetivo deste trabalho é avaliar a compreensibilidade das explicações de técnicas de IA explicável na perspectiva dos especialistas no domínio médico, por meio de entrevistas. E assim caracterizar a confiança e compreensão dos resultados produzidos por técnicas de IA explicável.

A questão central é analisar se a lógica revelada é aceitável pelo especialista no domínio e, se não, fornecer as recomendações apropriadas para melhorar essas técnicas.

1.4.2 Objetivos Secundários

A caracterização das técnicas de inteligência artificial explicável está vinculada aos seguintes objetivos secundários:

• Entender se os usuários especialistas no domínio médico confiam na inteligência artificial;

• Obter um conjunto de percepções, dificuldades, falhas e acertos e sugestões para a implantação da inteligência artificial explicável no domínio médico.

1.5 Relevância da Pesquisa

De acordo com Gunning [22] a eficácia dos sistemas de inteligência artificial, especialmente em aplicações críticas, como diagnóstico de doenças, negociação de ações e aplicações jurídicas, será limitada pela incapacidade das máquinas de explicar suas decisões e conclusões para os seres humanos. Assim, é importante construir uma inteligência artificial mais explicável, para que os humanos possam entender, confiar e gerenciar efetivamente os sistemas emergentes de inteligência artificial.

A necessidade de IA explicável é motivada principalmente por três razões [24] :

1. a necessidade de confiança;

2. a necessidade de interação;

3. a necessidade de transparência.

Embora existam diversos trabalhos relacionados a novas técnicas de sistemas de inteligência artificial explicável, poucos estudos avaliaram essas técnicas por uma perspectiva

do usuário. Foram encontrados poucos trabalhos que tenham realizado uma avaliação humana integrada em resultados de técnicas de sistemas de IA explicável populares (tais como: LIME [3, 6], SHAP [25–27] e Permutation Importance [28, 29] numa perspectiva do usuário especialista.

1.6 Escopo da Pesquisa

A pesquisa limita-se a analisar qualitativamente na perspectiva do usuário especialista no domínio médico (oncológico) as técnicas de IA explicável LIME, SHAP e Permutation Importance.

Nem todas as funcionalidades dessas técnicas foram consideradas. A avaliação ficou restrita aos gráficos de resultados das técnicas de inteligência artificial explicável (XAI).

1.7 Estrutura da Dissertação

A dissertação divide-se em oito capítulos, incluindo este capítulo de introdução:

O capítulo 2 (Fundamentação Teórica) apresenta conceitos sobre inteligência artificial, aprendizagem de máquina e seus respectivos algoritmos.

O capítulo 3 (Inteligência Artificial Explicável) apresenta informações sobre explicabilidade e interpretabilidade, inteligência artificial explicável, os tipos de técnicas e as respectivas técnicas.

O capítulo 4 (A Perspectiva do Usuário sobre as Técnicas de Inteligência Artificial Explicável) apresenta na perspectiva do usuário o desafio da avaliação das técnicas de inteligência artificial explicável, além de apresentar referências a outros trabalhos da literatura sobre a explicabilidade e avaliação das técnicas de inteligência artificial explicável.

O capítulo 5 (Metodologia de Pesquisa) apresenta os procedimentos utilizados durante o processo de pesquisa.

O capítulo 6 (Resultados da Avaliação Humana das Técnicas de Inteligência Artificial Explicável) detalha as avaliações realizadas com as entrevistas com os especialistas em oncologia e apresenta os resultados obtidos nas avaliações, bem como o cruzamento desses resultados, com objetivo de identificar as contribuições em comum e individuais para cada método de inteligência artificial explicável.

O capítulo 7 (Achados das Técnicas de Inteligência Artificial Explicável e Recomendações para o Desenvolvimento de Técnicas de Inteligência Artificial Explicável) apresenta os principais achados das técnicas de inteligência artificial explicável, uma das principais contribuições da dissertação, além de recomendações para a melhoria das técnicas de IA explicável.

O capítulo 8 (Conclusão) apresenta as conclusões da pesquisa.

2. Fundamentação Teórica

Este capítulo apresenta informações sobre inteligência artificial, paradigmas e subáreas dos sistemas de inteligência artificial. São apresentados também conceitos sobre aprendizagem de máquina e os diversos algoritmos que a compõem.

2.1 Inteligência Artificial

A Inteligência Artificial é um ramo da ciência da computação que simulam a capacidade humana de executar tarefas em ambientes complexos e de melhorar o desempenho aprendendo sem orientação constante de um usuário [8].

Segundo alguns autores, a inteligência artificial é dividida em dois paradigmas, simbólico e sub-simbólico:

• **Simbólico**: a inteligência artificial clássica (simbólica) é resolvida na estrutura pela chamada representação simbólica. Sua essência principal consiste em que, para determinados problemas elementares, temos processadores simbólicos disponíveis, que no local de entrada aceitam informações de entrada simbólicas e no local de saída oposto criam informações de saída simbólicas. O problema básico da inteligência artificial clássica inclui representação do conhecimento, processos de raciocínio, resolução de problemas, comunicação em linguagem natural, robótica, regras, ontologias e muito mais.

• **Sub-simbólico**: na teoria sub-simbólica (conexionista), as informações são processadas paralelamente por cálculos simples realizados pelos neurônios. Nesta abordagem, as informações são representadas por um simples pulso de sequência. Os modelos sub simbólicos são baseados em uma metáfora do cérebro humano, onde as atividades cognitivas do cérebro são interpretadas por conceitos teóricos que têm sua origem na neurociência. O problema básico da inteligência artificial sub-simbólica inclui aprendizado bayesiano, aprendizado profundo, conexionismo, redes neurais e muito mais.

2.1.1 Subáreas da Inteligência Artificial

Embora não haja um consenso de uma taxonomia em relação aos subcampos da inteligência artificial, seguem abaixo as mais comumente citadas na literatura [8]:

• **Planejamento e agendamento (Planning and scheduling)**: o planejamento e agendamento é um subcampo da IA dedicado à solução de problemas. Os problemas de planejamento e programação podem ser definidos como a escolha e seleção de ações que um determinado agente deve executar para que uma tarefa seja executada da melhor maneira possível, de maneira a atingir uma meta e maximizar o desempenho [8].

• **Processamento de linguagem natural**: é um subcampo da inteligência artificial focado em permitir que os computadores entendam e processem linguagens humanas, para aproximar os computadores de uma compreensão da linguagem em nível humano. Os computadores ainda não possuem o mesmo entendimento intuitivo da linguagem natural que os humanos. Eles não conseguem entender o que o idioma está realmente tentando dizer. Em poucas palavras, um computador não consegue ler nas entrelinhas [8].

• **IA e sociedade**: a inteligência artificial promete introduzir mudanças fundamentais em nossa sociedade, afetando tudo, desde negócios ao governo, vida profissional e vida pessoal. A IA e a sociedade são o subcampo que estuda os impactos da IA em nossa sociedade, incluindo economia, ética, formulação de políticas, filosofia, transparência, lei e justiça e muito mais.

• **Visão computacional**: a visão computacional é definida como "um subconjunto da inteligência artificial convencional que lida com a ciência de tornar computadores ou máquinas visualmente ativados, ou seja, eles podem analisar e entender uma imagem" [30] . A visão humana começa nos "olhos"da câmera biológica, que tira uma foto a cada 200 milissegundos, enquanto a visão por computador começa fornecendo informações à máquina. Isso torna o melhor caso para uma classe de algoritmos chamada Redes Neurais Convolucionais.

• **Robótica**: este subcampo tem como objetivo criar robôs, ou seja, agentes físicos que, por meio de sensores, atuadores e efetores, se tornam capazes de interferir no mundo real [8].

• **Representação do conhecimento e raciocínio**: é o subcampo da IA dedicado a representar fatos sobre o mundo de uma forma que um sistema de computador possa utilizar para resolver tarefas complexas do mundo real. Esse subcampo incorpora conclusões da filosofia e inclui o estudo das ontologias, que tem como objetivo organizar as coisas

do mundo em hierarquia de categorias e seus objetos, substâncias e medidas. A representação do conhecimento e raciocínio incorporam esses formalismos e ontologias para otimizar o processo de criação de sistemas de raciocínio (sistemas especializados) [8].

• **Sistemas baseados em agentes e multiagentes**: é um subcampo da inteligência artificial focado em estudar os agentes. Um agente é um ente que é capaz de perceber o ambiente em que está localizado e executar ações de modo a interagir nesse ambiente. Os agentes são considerados inteligentes quando suas e decisões são os melhores possíveis para os ambientes e situações em que foram inseridos. Um agente pode ser avaliado através de medidas de desempenho. Já os sistemas multiagentes são aqueles em que há mais de um agente no ambiente e esses agentes devem operar de maneira a maximizar seu desempenho. Os agentes podem operar em colaboração ou de maneira competitiva [8].

• **Incerteza na IA**: É um subcampo da IA relacionado aos agentes que podem precisar lidar com a incerteza, seja devido à observabilidade parcial, não determinismo ou uma combinação dos dois. Um agente pode nunca saber ao certo em que estado está ou onde será parar após uma sequência de ações.

• **Aprendizagem de máquina**: é um conjunto de métodos que permitem que os computadores aprendam com os dados para fazer e melhorar previsões (por exemplo, câncer). O aprendizado de máquina é uma mudança de paradigma da "programação normal", onde todas as instruções devem ser explicitamente fornecidas ao computador para a "programação indireta" que ocorre através do fornecimento de dados.

2.2 Aprendizagem de Máquina

O aprendizado de máquina (Machine Learning) é uma subárea da inteligência artificial. Se caracteriza como um conjunto de métodos que permitem aos computadores aprender com os dados para fazer e melhorar previsões. O aprendizado de máquina é uma mudança de paradigma da "programação normal", onde todas as instruções devem ser explicitamente fornecidas ao computador para a "programação indireta", que ocorre através do fornecimento de dados [18].

Uma outra definição de aprendizado de máquina é a utilização de algoritmos para extrair informações de dados brutos e representá-los através de um modelo matemático. Usamos então este modelo para fazer inferências (previsões) a partir de outros conjuntos de dados. E existem muitos algoritmos que permitem fazer isso, cabendo a um especialista escolher o algoritmo que melhor se encaixa em cada tipo de problema a resolver.

O principal objetivo do aprendizado de máquina é reproduzir nas máquinas o mesmo processo de aprendizagem dos seres humanos, através de algoritmos [8].

Algoritmo de aprendizado de máquina é o programa usado para aprender um modelo de aprendizado de máquina a partir de dados. Um algoritmo de aprendizado de máquina é um algoritmo capaz de aprender com dados. Mas o que queremos dizer com aprendizado? De acordo com Mitchell (1997) [31] que fornece uma definição sucinta: "um algoritmo aprende com a experiência E com relação a alguma classe de tarefas T e com a medida de desempenho P, se seu desempenho nas tarefas em T, medido por P, melhorar com a experiência E."

As tarefas (T) de aprendizado de máquina são geralmente descritas em termos de como o sistema de aprendizado de máquina deve processar um exemplo. Um exemplo é uma coleção de características que foram medidos quantitativamente a partir de algum objeto ou evento que queremos que o sistema de aprendizado de máquina processe. Normalmente, representamos um exemplo como um vetor $x \in \mathbb{R}^n$, em que cada entrada x_i do vetor é uma característica (feature).

Para avaliar as habilidades de um algoritmo de aprendizado de máquina, foi projetado uma medida quantitativa de seu desempenho. Normalmente, essa medida de desempenho P é específica da tarefa T sendo executada pelo sistema. Para tarefas como classificação e classificação com entradas ausentes, geralmente é medida a precisão (accuracy) do modelo. Precisão é apenas a proporção de exemplos para os quais o modelo produz a saída correta.

2.2.1 Aprendizagem de máquina supervisionado e não supervisionado

As técnicas de aprendizado de máquina empregam um princípio de inferência denominado indução, no qual obtém-se conclusões genéricas a partir de um conjunto de dados de treinamento. Pode ser dividido em dois tipos principais: supervisionado e não-supervisionado.

Aprendizado supervisionado: No aprendizado supervisionado tem-se a figura de um professor, o qual apresenta o conhecimento do ambiente externo através de um conjunto de dados (exemplos) na forma: dados de entrada, dados da saída desejada. O algoritmo de aprendizado de máquina extrai a representação do conhecimento a partir desses dados de treinamento. o aprendizado supervisionado envolve a observação de vários exemplos de um vetor aleatório x e um valor ou vetor associado, aprendendo a predicar a partir de x, geralmente estimando $p(y|x)$ [32].

Aprendizado Não Supervisionado: O aprendizado não supervisado ocorre quando o algoritmo pode encontrar padrões e relações em um conjunto de dados. O objetivo de um algoritmo de aprendizado não supervisionado é organizar os dados agrupando-os em grupos de exemplos relacionados ou descrever sua estrutura. Este tipo de aprendizado assemelha-se aos métodos que nós, seres humanos, usamos para classificar as coisas. São utilizados quando o objetivo for encontrar padrões ou tendências que possam ajudar no entendimento dos dados.

O aprendizado não supervisionado envolve a observação de vários exemplos de um vetor aleatório x e a tentativa de aprender implícita ou explicitamente a distribuição de probabilidade $p(x)$ ou algumas propriedades interessantes dessa distribuição [32].

A distinção entre algoritmos supervisionados e não supervisionados não é formal e rigidamente definida, porque não há um teste objetivo para distinguir se um valor é uma característica ou um destino fornecido por um supervisor.

Aprendizagem semi-supervisionada: A aprendizagem semi-supervisionada é parcialmente supervisionada e parcialmente não supervisionada. Nesse tipo de aprendizagem, alguns exemplos incluem um objetivo de supervisão, mas outros não. No aprendizado em várias instâncias, uma coleção inteira de exemplos é rotulada como contendo ou não contendo um exemplo de classe, mas os membros individuais da coleção não são rotulados [32].

Aprendizado por Reforço (Reinforcement Learning): No aprendizado por reforço, o algoritmo escolhe uma ação em resposta a cada ponto de dados, ou seja, o aprendizado ocorre por tentativa e erro ou força bruta. O aprendizado por reforço é comum em robótica, em que o conjunto de leituras do sensor, em um ponto no tempo, é um ponto de dados e o algoritmo deve escolher a próxima ação do robô.

Teoria do Aprendizado Estatístico é a teoria por trás de vários algoritmos de aprendizagem de máquina em aprendizagem supervisionada. Sendo f um classificador e F o conjunto de todos os classificadores gerado por um algoritmo de aprendizagem de máquina. No processo de aprendizado, esse algoritmo utiliza um conjunto de treinamento T, composto de n pares (x_i, y_i), para gerar um classificador particular $\hat{f} \in F$. O objetivo do processo de aprendizado é encontrar um classificador que separe os dados das classes.

Um conceito importante empregado em aprendizagem de máquina é o de generalização de um classificador, definida como a sua capacidade de prever corretamente a classe de novos dados.

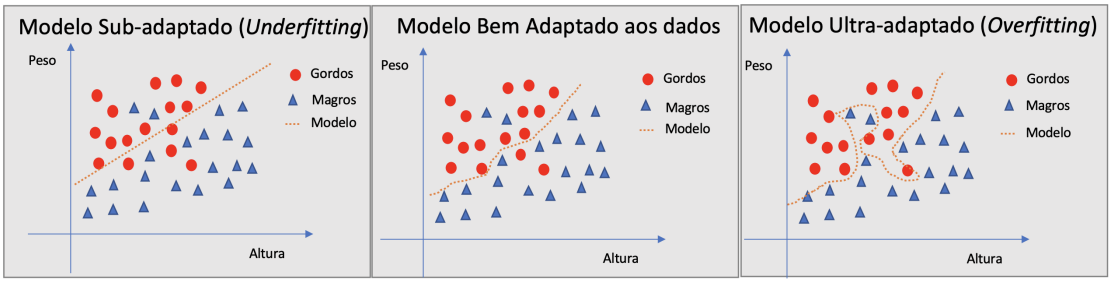

Figura 2 – Modelos gerados por técnicas de Machine learning com mesma base de dados: os riscos de under and overfitting. Geração de um classificador do estado físico de clientes de um nutricionista. Triângulos representam pessoas magras e círculos pessoas gordinhas.

Na 1º imagem da figura 2, tem se a ocorrência de um subajustamento (underfitting). Um subajustamento ocorre quando o modelo não é capaz de obter um valor de erro suficientemente baixo no conjunto de treinamento.

Na 2º imagem da figura 2, temos um exemplo de capacidade apropriada (appropriate capacity). A capacidade apropriada ocorre quando o modelo classifica corretamente grande parte dos dados, sem se fixar demasiadamente em qualquer ponto individual.

Na 3º imagem da figura 2, tem se a ocorrência de um superajustamento (overfitting). O superajustamento ocorre quando a diferença entre a taxa de erro de treinamento e a taxa de erro de teste é muito grande, ou seja, o modelo se especializa nos dados utilizados em seu treinamento.

A Teoria de Aprendizado Estatístico estabelece condições matemáticas que auxiliam na escolha de um classificador particular $\hat{f}$ a partir de um conjunto de dados de treinamento, com o objetivo de obter um desempenho otimizado para os novos dados. Essas condições incluem o desempenho do classificador no conjunto de treinamento e a sua complexidade. Na aplicação da Teoria de Aprendizado Estatístico, assume-se inicialmente que os dados do domínio em que o aprendizado está ocorrendo são gerados de forma independente e identicamente distribuída de acordo com uma distribuição de probabilidade $P(x, y)$, que descreve a relação entre os dados e os seus rótulos [32].

2.3 Principais Algoritmos de Aprendizagem Supervisionada

Algoritmos de aprendizado supervisionado são algoritmos que aprendem a associar alguma entrada a alguma saída, dado um conjunto atraente de exemplos de entradas x e saídas y. Em muitos casos, as saídas y podem ser difíceis de coletar automaticamente e

devem ser fornecidas por um “supervisor” humano, mas o termo ainda se aplica mesmo quando as metas do conjunto de treinamento foram coletadas automaticamente.

2.3.1 Naive Bayes

O classificador Naïve Bayes é um método probabilístico que considera a probabilidade condicional de um evento aocntecer uma vez que um conjunto de outros eventos acontecem. Por exemplo, qual a probabilidade de um paciente internado estar com COVID uma vez que apresenta febre alta e ter perdido o olfato, mas que vinha fazendo isolamento social? Conforme apresentado na Eq. 2.1, a probabilidade de um evento A ocorrer dado que um evento B já ocorreu pode ser calculado como a probabilidade de haver o sintoma dado que há o sintoma (obtido pelo histórico sobre a doença) vezes a prior sobre a doença, sito é , a probabilidade da doença naquela população, dividido pela probabilidade de haver o sintoma (também registrado na base histórica) [1]. Classificadores bayesianos usam esta formula para criar diagnósticos e previsões usando cálculos dessa natureza.

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{2.1}$$

Ele é baseado na suposição simplificadora de que os valores dos atributos são condicionalmente independentes dado o valor alvo.

2.3.2 K-Nearest Neighbours (KNN)

O KNN é um método baseado em instância. Objetos relacionados ao mesmo conceito são semelhantes entre si. A Regra do KNN é classificar xz atribuindo a ele o rótulo representado mais frequentemente dentre as k amostras mais próximas e utilizando um esquema de votação [1].

2.3.3 Árvore de decisão (Decision Tree)

A árvore de decisão representa uma função que assume como entrada um vetor de valores de atributo e retorna uma "decisão- um único valor de saída. Os valores de entrada e saída podem ser discretos ou contínuos. A classificação das instâncias é realizada ordenando as árvores a partir da raiz até alguma folha. Outra forma de representação das árvores de decisão

Árvores de decisão classificam instâncias ordenando as árvores acima (ou abaixo), a partir da raiz até alguma folha. Árvores de decisão também podem ser representadas como

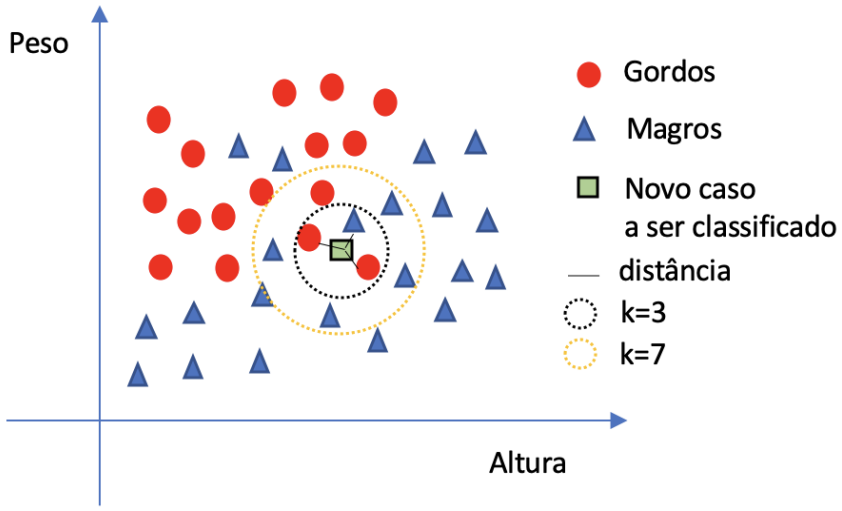

Figura 3 – Exemplo de funcionamento de regras de KNN

conjuntos de regras SE–ENTÃO-SENÃO (IF–THEN-ELSE) [1]. No capítulo Introdução desta dissertação apresentamos um exemplo de árvore de decisão, vide Figura 1.

2.3.4 Máquinas de Vetores de Suporte (Support Vector Machines)

O algoritmo de máquina de vetores de suporte (SVM – Support Vector Machines) é um método não paramétrico de aprendizagem supervisionado, usado para estimar uma função que classifica os dados de entrada em duas ou mais classes. Hiperplanos são um separador de margem máxima, ou seja, um limite de decisão com a maior distância possível aos pontos de exemplo, sendo que sua dimensão está condicionada ao número de características.

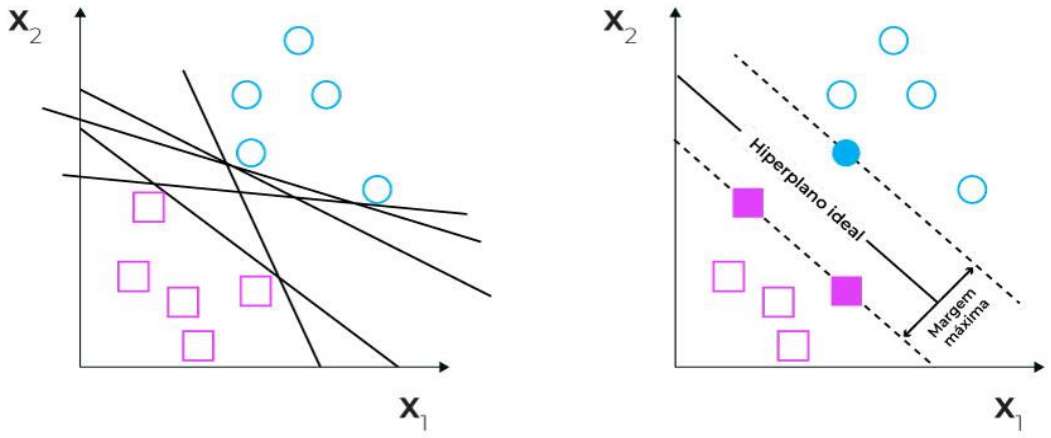

Figura 4 – SVM. Figura retirada de [1].

O objetivo do algoritmo é encontrar um hiperplano em um espaço $N-dimensional$ (N - o número de características) que classifica distintamente os pontos de dados, conforme exibido na figura 4. Existe uma grande variedade de hiperplanos que podem ser escolhidos para separar as duas classes de pontos de dados, sendo que o objetivo principal é que o modelo generalize bem. Para isso é construído um hiperplano com a margem máxima, ou seja, um limite com maior distância possível entre os pontos de dados das duas classes. Geralmente, os dados que não são linearmente separáveis no espaço de entrada original são facilmente separáveis no espaço de maior dimensão [1, 8].

2.3.5 Métodos ensemble

Métodos Ensemble são uma categoria de algoritmos de aprendizagem de máquina, que podem ser usados tanto em aprendizagem supervisionada quanto não supervisionada. Foi criado por Tom Mitchel [33], quando fez um estudo com 23 algoritmos diferentes e construindo uma árvore de decisão para prever o melhor algoritmo a ser usado, dadas as propriedades de um conjunto de dados. Construir um ensemble consiste em dois passos: construir diversos modelos e combinar as suas respectivas estimativas. A combinação pode ser feita por meio de votação ou por meio de pesos das estimativas dos modelos. Os Métodos Ensemble permitem aumentar consideravelmente o nível de precisão nas suas previsões. [1, 8].

Os principais algoritmos de métodos ensemble são: Random Forest, Bagging, Adaboost, Gradient Boosting.

Random Forest é um método que combina diversas árvores de decisão e adiciona um componente estocástico para criar mais diversidade entre as árvores de decisão.

Bagging ou Boostrap Aggregation aplica técnicas de boostrap (reamostragem) no conjunto de dados de treino a fim de construir diversas árvores de decisão e depois obtém a melhor estimativa por votação entre as árvores ou pelo peso das estimativas.

Adaboost constói diversos modelos de forma iterativa, variando o peso nos exemplos no conjunto de dados de treino, penalizando exemplos com muitos erros e reduzindo o peso daqueles com estimativa incorreta ou menos precisa.

Gradient Boosting é uma extensão do AdaBoost com uma variedade de funções de erro para classificação e regressão.

Conforme dito acima, o algoritmo randon forest e outros algoritmos do método ensemble combinam diferentes árvores de decisão e/ou outros algoritmos para obter uma

previsão agregada. Embora esses algoritmos sejam eficazes contra o sobreajuste, a combinação de modelos torna a interpretação do conjunto geral mais complexa do que cada um de seus aprendizes de árvores compostas, tornando-os algoritmos de caixa-preta. Dessa maneira para o usuário interpretar os resultados, deve recorrer a técnicas de explicabilidade post-hoc [7].

2.4 Principais Algoritmos de Aprendizagem Não Supervisionada

Os principais algoritmos não supervisionados são o K-Means, o Principal Component Analysis (PCA) e o Singular Value Decomposition (SVD).

2.4.1 Principal Components Analysis (PCA)

O PCA é um algoritmo de aprendizado não supervisionado, onde cada componente principal é uma combinação de atributos presentes no conjunto de de dados. É um algoritmo que aprende uma representação de dados que é baseada em dois dos critérios para uma representação simples. O PCA aprende uma representação que possui menor dimensionalidade que a entrada original e também aprende uma representação cujos elementos não têm correlação linear entre si. O PCA é associado à ideia de redução de massa de dados, com menor perda possível da informação.

Este é o primeiro passo em direção ao critério de aprender representações cujos elementos são estatisticamente independentes. Para alcançar total independência, um algoritmo de aprendizado de representação também deve remover os relacionamentos não lineares entre variáveis.

O PCA aprende uma transformação linear ortogonal dos dados que projetam uma entrada x para uma representação z. O PCA é um método simples e eficaz de redução de dimensionalidade que preserva o máximo possível de informações nos dados (medido pelo erro de reconstrução de mínimos quadrados) [32].

2.4.2 k-means clustering

O algoritmo de agrupamento k-means divide o conjunto de treinamento em k diferentes agrupamentos de exemplos que estão próximos um do outro. Podemos, assim, pensar no algoritmo como um vetor de código unidimensional k que representa uma entrada x. Se x pertence ao cluster i, então $h_i = 1$ e todas as outras entradas da representação h são zero.

O código one-hot fornecido pelo cluster de k-means é um exemplo de representação esparsa, porque a maioria de suas entradas é zero para cada entrada. Posteriormente, foram desenvolvidos outros algoritmos que aprendem representações esparsas mais flexíveis, nas quais mais de uma entrada pode ser diferente de zero para cada entrada x. Os códigos one-hot são um exemplo extremo de representações esparsas que perdem muitos dos benefícios de uma representação distribuída. O código one-hot ainda confere algumas vantagens estatísticas (naturalmente transmite a ideia de que todos os exemplos no mesmo cluster são semelhantes entre si) e confere a vantagem computacional de que toda a representação pode ser capturada por um único inteiro [32].

O algoritmo k-means funciona inicializando k diferentes centroides $\{\mu(1), \ldots, \mu(k)\}$ para diferentes valores, alternando entre duas etapas diferentes até a convergência. Em uma etapa, cada exemplo de treinamento é atribuído ao cluster i, onde i é o índice do centroide mais próximo $\mu(i)$. Na outra etapa, cada centroide $\mu(i)$ é atualizado para a média de todos os exemplos de treinamento $x(j)$ atribuídos ao cluster i [32].

2.5 Redes Neurais Artificiais

Redes Neurais Artificiais são um um paradigma da inteligência artificial. Uma rede neural usa uma rede de neurônios artificiais para resolver problemas de aprendizagem, basicamente imitando o processo de aprendizagem humano. A rede que você vê na figura 5 é uma rede neural artificial composta de neurônios interligados, onde a primeira camada da rede (camada de entrada) é representado por $I_1, I_2, I_3, ..., I_{256}$, a última camada (camada de saída) é representada por $O_1, ..., O_{16}$ e a camada entre as duas é referida como camada oculta, é representado por $H_1, ..., H_{16}$.

As Redes Neurais Artificiais são um tipo de algoritmo de aprendizagem de máquina que podem ser aplicados a quase todas as tarefas de aprendizagem de máquina. As redes neurais artificiais são melhor aplicadas a problemas onde os dados de entrada e os dados de saída são bem definidos ou, pelo menos, bastante simples, mas o processo que relaciona a entrada com a saída é extremamente complexo.

O neurônio biológico

A figura 6 exibe um modelo matemático simplificado de um neurônio estático. A Fórmula do Neurônio Artificial 2.2 e a Fórmula da Função de Ativação 2.3.

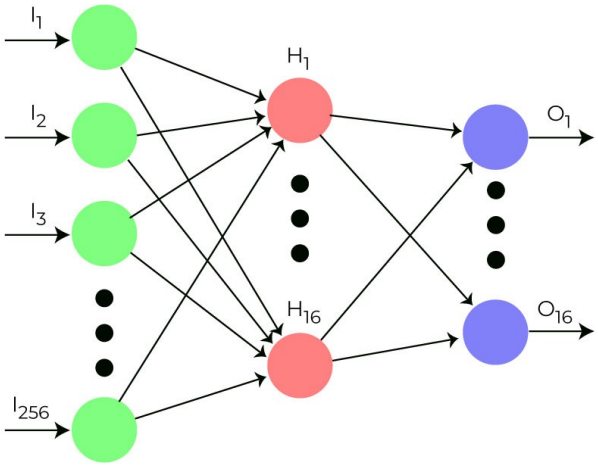

Figura 5 – Rede neural artificial composta de neurônios interligados. Figura retirada de [2].

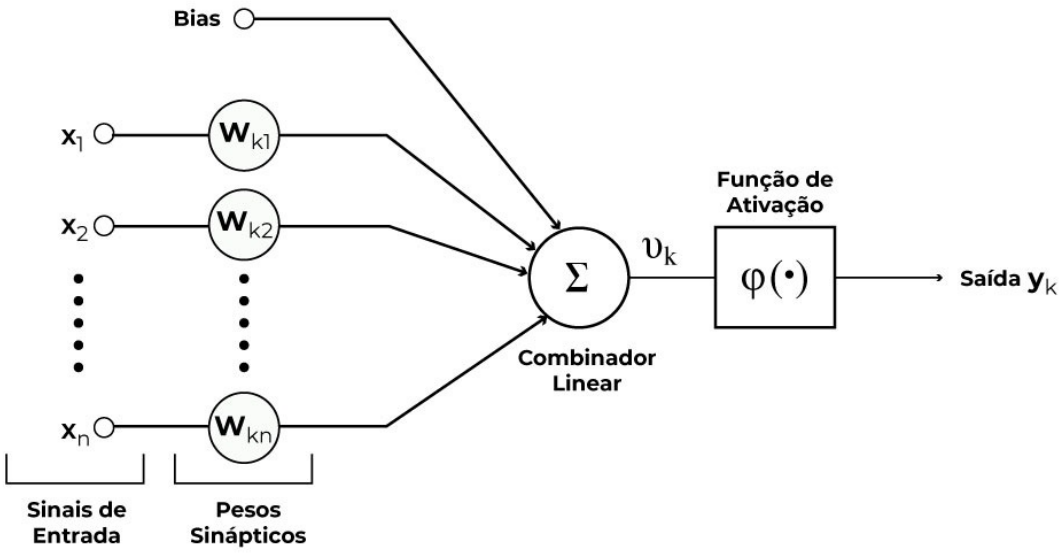

Figura 6 – Modelo matemático simplificado de um neurônio. Figura retirada de [2].

$$u_k = \sum_{j=1}^{m} w_{kj} * x_j \tag{2.2}$$

$$y_k = \varphi(u_k) \tag{2.3}$$

O modelo neuronal matemático também pode incluir uma polarização ou bias de entrada ao somatório da função de ativação e consequentemente aumenta a capacidade de aproximação da rede.

O Perceptron

Um perceptron funciona analogamente a um neurônio. Os perceptrons são capazes de aprender conectado, com o erro. O principal objetivo do Perceptron é aprender iterativamente os pesos sinápticos de tal forma que a unidade de saída produza a saída correta para cada exemplo [2].

As principais limitações do Perceptron são:

- Um único Perceptron consegue resolver somente funções linearmente separáveis;

- O perceptron não consegue gerar um hiperplano para separar os dados em funções não linearmente separáveis.

As limitações do Perceptron inspiraram os modelos mais avançados de redes neurais.

Adaline (Adaptive Linear Element ou ADAptive LInear NEuron)

O Adaline é similar ao Perceptron, sendo a principal diferença o seu algoritmo de treinamento. O Perceptron ajusta os pesos somente quando um padrão é classificado incorretamente, já o Adaline utiliza a Regra Delta para minimizar o erro médio (MSE) após cada padrão ser apresentado,ou seja, os pesos são ajustados proporcionalmente ao erro [2].

$$E = \frac{1}{2M} \sum_{p=1}^{M} (d^p - S^p)^2 = \frac{1}{2M} \sum_{p=1}^{M} [d^p - \sum_{i=0}^{N} \omega_i x_i^p]^2 \tag{2.4}$$

Redes Multilayer Perceptron

Redes Multilayer Perceptron são redes diretas (feed forward) que possuem mais de uma camada de neurônios entre as camadas de entrada e saída, chamada de camada oculta. Esta camada adiciona um poder maior em relação às redes Perceptron de camada única, que classifica apenas padrões linearmente separáveis, sendo os neurônios ocultos responsáveis por capturar a não-linearidade dos dados [2]. Isso se deve pela possibilidade de transformação do vetor de entradas para se adaptar a comportamentos de não linearidade e descontinuidade.

2.5.1 Redes Neurais Profundas (Deep Learning)

Redes Neurais Artificiais Profundas ou Aprendizagem Profunda é uma subárea do aprendizagem de máquina e é uma evolução das redes neurais, ou seja, é um tipo de rede

neural artificial. Usam algoritmos para processar dados e imitar o processamento feito pelo cérebro humano. A figura 7 exibe uma rede neural profunda de reconhecimento de imagens com a camada de entrada (imagens), com 3 (três) camadas ocultas (arestas, combinação de arestas e modelos de objetos) e a camada de saída com a imagem "reconhecida".

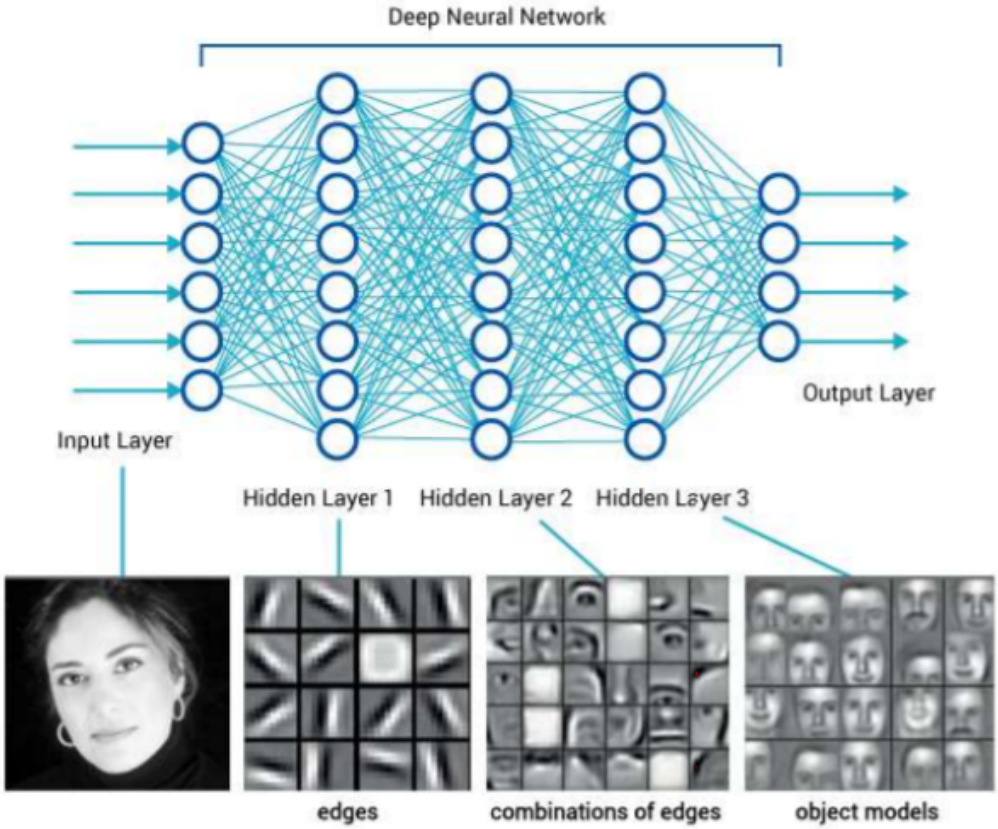

Figura 7 – Rede Neural Profunda. Figura retirada de [2].

O processo de aprendizagem dos algoritmos de Aprendizagem Profunda se baseia no processo de aprendizagem do nosso cérebro, principalmente na parte conhecida como córtex visual, o chamado processo de Aprendizagem Hierárquica.

Aprendizagem Profunda usa camadas de neurônios matemáticos para processar dados, compreender a fala humana e reconhecer objetos visualmente. A informação é passada através de cada camada, com a saída da camada anterior fornecendo entrada para a próxima camada.

A primeira camada em uma rede é chamada de camada de entrada e a última de camada de saída. Todas as camadas entre as duas são referidas como camadas ocultas, sendo cada camada um algoritmo simples e uniforme contendo um tipo de função de ativação. O uso de várias camadas ocultas permite uma acumulação mais sofisticada de elementos simples a outros mais complexos. São considerados dois aspectos de complexidade da

arquitetura de um modelo: Número de neurônios por camada e Número de camadas [2].

As principais arquiteturas de redes são Redes Neurais Convolucionais e Redes Neurais Recorrentes:

Redes Neurais Convolucionais (Convolutional Neural Networks, CNNs)

São redes neurais artificiais profundas que aplica uma operação matemática denominada convolução em pelo menos uma das camadas, ao invés de multiplicação geral de matrizes.Convolução é um tipo especializado de operação linear. Redes convolucionais são redes neurais simples são muito utilizadas para classificação de imagens, agrupamento por similaridade (busca de fotos), reconhecimento de objetos e reconhecimento óptico de caracteres (OCR) para digitalizar texto. O reconhecimento do aprendizado profundo deve muito à eficácia das redes convolucionais no reconhecimento de imagens [2, 32].

Redes Neurais Recorrentes (Recurrent Neural Networks, RNN's

As Redes Neurais Recorrentes são um conjunto de algoritmos especialmente úteis para o processamento de dados sequenciais, como som, áudio, dados de séries temporais ou linguagem natural. Redes neurais recorrentes constituem uma ampla classe de redes cuja evolução do estado depende tanto da entrada corrente quanto do estado atual. A figura **??** mostra a ideia por trás das redes neurais recorrentes, que é fazer uso de informações sequenciais. Basicamente, o objetivo em usar redes neurais recorrentes é examinar os sistemas reais e seus comportamentos ao longo do tempo em resposta aos estímulos [2, 32].

3. Inteligência Artificial Explicável

A Inteligência Artificial Explicável (XAI) visa revelar o raciocínio oculto dos agentes inteligentes. O principal objetivo das técnicas de Inteligência Artificial Explicável é apoiar a compreensão humana, ou seja, tornar os sistemas de IA mais transparentes, interpretativos e explicáveis [9]. Portanto, os resultados dos algoritmos de aprendizagem de máquina podem divulgar a explicação causal e os passos dados nas saídas da máquina, previsões e recomendações.

Como resultado, especialistas no domínio, pesquisadores e demais usuários podem finalmente entender como e por que um algoritmo tomou uma determinada decisão.

Portanto, o objetivo principal das pesquisas recentes em inteligência artificial explicável é alcançar explicabilidade e apoiar a compreensão humana.

Este capítulo apresenta informações sobre interpretabilidade e explicabilidade, bem como os critérios para categorizar as técnicas de IA explicável.

3.1 Conceitualização

3.1.1 O que é interpretabilidade e explicabilidade

A nomenclatura mais comumente usada nas comunidades de XAI e áreas afins são interpretabilidade e explicabilidade.

A interpretabilidade é definido como a capacidade de explicar ou fornecer um significado de maneira compreensível para os humanos. Já a explicabilidade está relacionado à noção de explicação como uma interface entre humanos e um tomador de decisão que é, ao mesmo tempo, precisa ao tomador de decisão e compreensível para os seres humanos [34].

O que geralmente chamamos de interpretabilidade e explicabilidade, pode ser definido como o quanto bem um ser humano pode entender as decisões num determinado contexto. Não há definição matemática de interpretabilidade e explicabilidade.

Lipton [35] afirma que explicabilidade é uma noção contextual e não absoluta. Em seu trabalho ele busca identificar propriedades desejáveis para sistemas interpretáveis, com destaque para transparência, confiança e interpretabilidade post-hoc. Esse último relacionada à capacidade do sistema de oferecer informações úteis sobre seus resultados para os diversos perfis de usuário. Interpretabilidade é o grau em que um humano pode prever consistentemente o resultado do modelo.

O objetivo principal da explicabilidade é a atribuição de eventos causais que permitem ao usuário responder a perguntas do "por que" [17].

Uma explicação semântica é criada para entender o comportamento dos sistemas de IA. Como os humanos preferem perguntas contrastantes [14], é relevante notar que a pergunta "por que"se torna mais desafiadora.

As pessoas não perguntam por que o evento Q aconteceu, mas querem saber por que o evento P não aconteceu. Isso adiciona um raciocínio mais sofisticado às técnicas de IA explicável, pois o processo deve considerar o raciocínio contrafactual para concentrar-se não apenas nos eventos que aconteceram, mas também simular ocorrências que não aconteceram.

Ao comparar modelos de aprendizagem de máquina, além de desempenho, é dito que um modelo tem uma melhor interpretabilidade do que outro modelo se os resultados gerados a partir do modelo são mais fáceis de serem entendidas por um ser humano do que as de outro modelo [35].

Objetivo das técnicas de IA Explicável (XAI)

Qualquer modelo de aprendizado de máquina possui uma função de resposta que tenta mapear e explicar relacionamentos e padrões entre as variáveis independentes (de entrada) e as variáveis dependentes (de destino ou resposta).

Um dos objetivos das técnicas de XAI é tentar desvendar os motivos e a lógica interna das tecnologias de aprendizado de máquina, especialmente algoritmos de caixa-preta, como redes neurais, máquinas de vetores de suporte, floresta aleatória e outros.

Modelos explicáveis promovem o diálogo entre as áreas de interação homem-computador (HCI) e inteligência artificial.

3.2 Critérios para categorizar os métodos de interpretação de modelos

Segundo Molnar [18] existem critérios específicos que podem ser usados para categorizar os métodos de interpretação do modelo (muitos dos quais serão apresentados neste capítulo).

3.2.1 Intrínseco versus Post-hoc

Este critério distingue se a interpretabilidade é alcançada restringindo a complexidade do modelo de aprendizado de máquina (intrínseco) ou aplicando métodos que analisam o modelo após o treinamento (post hoc) [18].

A interpretabilidade intrínseca refere-se a algoritmos de aprendizado de máquina que são considerados interpretáveis devido à sua estrutura simples (como regressão linear, regressão logística, modelos baseados em árvore, k-nearest neighbors, aprendizado baseado em regras, modelos aditivos gerais e modelos bayesianos) [7].

A interpretabilidade post hoc refere-se à aplicação de métodos de interpretação após o treinamento do modelo, ou seja, um método separado (LIME, importância das características, gráficos de dependência parcial) deve ser aplicado ao modelo para explicar suas decisões. Esses métodos foram criados para explicar as decisões dos algoritmos de caixa-preta, como: métodos ensemble, SVM, redes neurais multicamadas, redes neurais profundas (redes neurais convolucionais e redes neurais recorrentes). Embora a motivação da criação dos métodos post hoc fosse a explicação de algoritmos de caixa-preta, os métodos post hoc também podem ser aplicados a modelos intrinsecamente interpretáveis [7, 18].

3.2.2 Dependente de Modelo versus Independente de Modelo (Model Agnostic)

As técnicas dependente de modelo são aquelas aplicáveis a um algoritmo de aprendizado de máquina específico. Incluem os modelos interpretáveis e métodos específicos do modelo:

Modelos interpretáveis são algoritmos que transmitem algum grau de interpretabilidade por si mesmos, ou seja, um modelo é considerado interpretável se, por si só, for compreensível.

Métodos específicos de modelo são técnicas post hoc (aplicação de métodos de interpretação após o treinamento do modelo) que analisam um algoritmo de aprendizagem específico, como redes neurais profundas, por exemplo. A desvantagem dos métodos es-

pecíficos de modelo é o fato de estar vinculado a um tipo de modelo específico e será impossível mudar para outro [18].

As técnicas independentes de modelo (model-agnostic) podem ser usadas em qualquer modelo de aprendizado de máquina, ou seja, permite flexibilidade na escolha de modelos. Por definição, esses métodos não têm acesso a nenhuma propriedade interna do modelo, como pesos, restrições ou suposições [18].

3.2.3 Local versus Global

Essa classificação de interpretação de modelos avalia se o método de interpretação explica uma única previsão ou todo o comportamento do modelo.

Como definimos o escopo e os limites da interpretabilidade? Alguns aspectos úteis podem ser a transparência, justiça e responsabilidade de um modelo. Interpretações globais e locais do modelo são maneiras claras de definir o escopo da interpretação de um modelo [35].

A figura 8 mostra a diferença entre Interpretação Global e Interpretação Local. Interpretação global explica o comportamento do modelo e interpretação local explica uma única previsão.

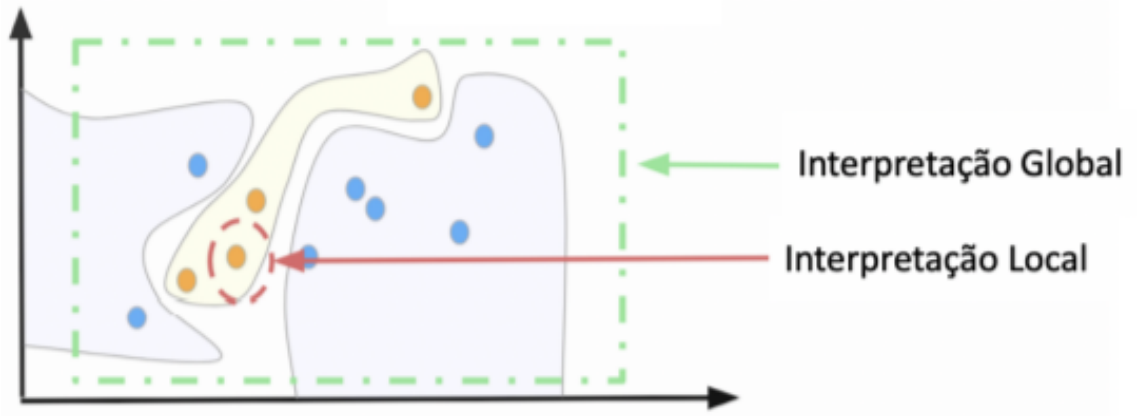

Figura 8 – Interpretação Global versus Interpretação Local. Figura retirada de [3].

Interpretação Global

Trata-se de tentar entender "Como o modelo faz previsões?"e “Como os subconjuntos do modelo influenciam as decisões do modelo?”. Para compreender e interpretar todo o modelo de uma só vez, precisamos de interpretabilidade global. Sistemas que focam na interpretabilidade global geram explicações das decisões do sistema com base nas interações condicionais entre as variáveis dependentes (resposta) e as características/variáveis

independentes (preditores) no conjunto de dados completo.

Para explicar a saída do modelo global, você precisa do modelo treinado, do conhecimento do algoritmo e dos dados. A explicação gerada transmite uma visão holística de como o sistema mapeou características (features) aos resultados e como cada um dos componentes aprendidos, como pesos, outros parâmetros e estruturas, foram gerados [18].

Essa abordagem considera que mostrar as interações entre as características e a importância relativa entre elas formam a base da explicação, ou seja, entender quais características usadas no modelo foram mais influentes para a decisão é sempre um passo para entender a interpretação global. Entretanto, visualizar a relação entre características após mais de duas ou três dimensões é bastante difícil para o ser humano. Por isso, é comum que as explicações agrupem observações em conjuntos e subconjuntos de características, que podem influenciar as previsões do modelo em um conhecimento global. É necessário um conhecimento completo da estrutura do modelo, suposições e restrições para uma interpretação global.

Interpretação Local

Trata-se de tentar entender as razões do modelo tomar decisões específicas para uma instância em particular . Para a interpretabilidade local, a estrutura global e as suposições inerentes a um modelo como um todo não são importantes. Para entender a previsão para uma dada instância ou caso, foca-se especificamente nas características da instância de dados. A análise é feita considerando apenas as caraterísticas e cenários que afetam a instância [3, 18].

Com isso as explicações são mais ricas para se entender o que se passou com aquele dado. O risco é de se ter um overfitting na explicação. As distribuições de dados locais e os espaços de recursos podem se comportar completamente diferentes e fornecer explicações mais precisas, em oposição às interpretações globais. O framework Local Interpretable Model-Agnostic Explanation (LIME) [3] é um dos métodos mais utilizados na abordagem local e que pode ser usado para gerar explicações locais independente do modelo aprendizagem de máquina utilizado.

3.3 Técnicas de inteligência artificial explicável

Molnar e Christoph [18], em seu livro "Interpretable Machine Learning - A Guide for Making Explainable Black Box Models", sugere 4 tipos diferentes de técnicas XAI: métodos interpretáveis, métodos específicos de modelo, métodos independentes (agnósticos)

de modelo e métodos de explicação baseado em exemplos.

3.3.1 Métodos interpretáveis (Interpretable Models)

O oposto de uma caixa preta às vezes é chamado de caixa branca e é referido na literatura como modelo interpretável. O aprendizado de máquina interpretável refere-se a métodos e modelos que tornam o comportamento e as previsões dos sistemas de aprendizado de máquina compreensíveis para os seres humanos.

Modelos interpretáveis significa que o algoritmo de aprendizagem de máquina são por si só interpretáveis, ou seja, os mecanismos internos de decisão são compreensíveis e transparentes e o algoritmo transmitem algum grau de interpretabilidade para os usuários [7,36]. Os principais algoritmos interpretáveis de acordo com alguns autores são [36–39]: regressão linear, regressão logística, árvores de decisão, k-nearest neighbors, modelos bayesianos, aprendizado baseado em regras, modelo aditivos gerais (GAM) e modelos linear generalizado (GLM).

3.3.2 Métodos dependentes de modelo

Métodos dependentes de modelo são técnicas post hoc que foram criadas para explicar a previsão de um algoritmo de aprendizagem de máquina de caixa preta específico como: redes neurais profundas e máquinas de vetores de suporte (SVM), por exemplo.

O foco desse trabalho foi os métodos dependentes de modelo baseados em redes neurais profundas (Deep Neural Networks). Veja abaixo os principais métodos baseados em redes neurais profundas:

• **iNNvestigate**. É uma biblioteca que facilita a análise de previsões de redes neurais e a comparação de diferentes métodos de análise. Isso é feito fornecendo um interface e implementações para muitos métodos de análise, além de disponibilizar ferramentas para treinamento e comparação de métodos. Em particular, ele contém implementações de referência para muitos métodos (PatternNet, PatternAttribution, LRP) e aplicativos de exemplo para um grande número de aplicativos de última geração. O objetivo é que a biblioteca ofereça suporte ao campo de análise de aprendizado de máquina e facilite a pesquisa usando redes neurais em domínios como design de medicamentos ou análise de imagens médicas [40].

• **SVCCA (Singular Vector Canonical Correlation Analysis)**. É uma técnica e ferramenta para comparar rapidamente duas representações, de uma maneira que é invariável

à transformação afim (permitindo a comparação entre diferentes camadas e redes) e rápido na computação (permitindo que mais comparações sejam calculadas do que com os métodos anteriores). Essa ferramenta foi implementada para medir a dimensionalidade intrínseca das camadas, mostrando em alguns casos uma parametrização desnecessária; para sondar a dinâmica de aprendizado durante o treinamento, descobrindo que as redes convergem para representações finais de baixo para cima; para mostrar onde as informações específicas da classe redes são formadas; e sugerir novos regimes de treinamento que economizem simultaneamente a computação e se ajustem menos [41].

• **Grad-CAM (Gradient-weighted Class Activation Mapping)**. É uma técnica para produzir "explicações visuais"para decisões de uma grande classe de modelos baseados em redes neurais convolucionais (CNN), tornando-os mais transparentes. Essa abordagem usa os gradientes de qualquer conceito de destino, fluindo para a camada convolucional final para produzir um mapa de localização aproximado, destacando as regiões importantes da imagem para prever o conceito. A figura 9 apresenta uma visão geral da técnica Grad-CAM. Diferentemente das abordagens anteriores, o Grad-CAM é aplicável a uma ampla variedade de famílias de modelos de CNN: (1) CNNs com camadas totalmente conectadas (por exemplo, VGG), (2) CNNs usadas para saídas estruturadas (por exemplo, legendagem), (3) CNNs usado em tarefas com entradas multimodais ou aprendizagem por reforço, sem alterações arquitetônicas ou treinamento.

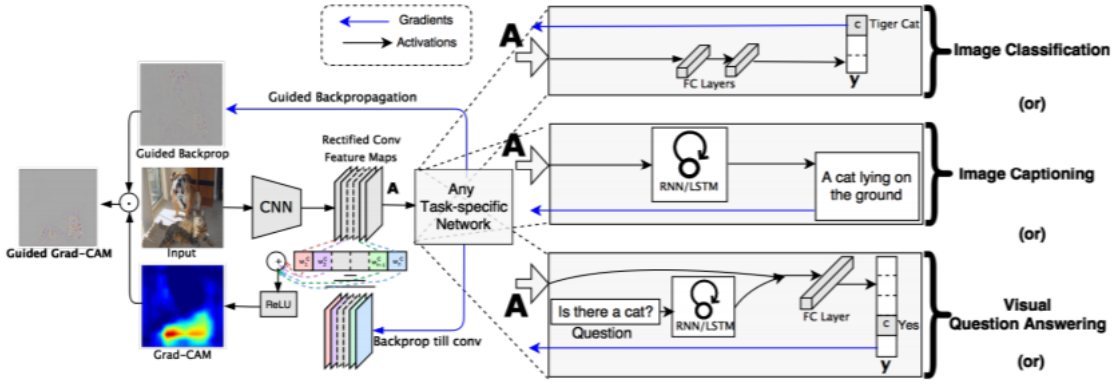

Figura 9 – Grad-CAM. Figura retirada de [4].

O Grad-CAM foi combinado com visualizações refinadas existentes para criar uma visualização discriminativa de classe de alta resolução para ser aplicada aos modelos de classificação de imagens, legendas de imagens e resposta visual a perguntas (VQA), incluindo arquiteturas baseadas em ResNet [4].

• **DeepLIFT (Deep Learning Important FeaTures)**. É um método para decompor a previsão de saída de uma rede neural em uma entrada específica, retropropagando as contribuições de todos os neurônios da rede para todos os recursos da entrada. O DeepLIFT

compara a ativação de cada neurônio à sua 'ativação de referência' e atribui pontuações de contribuição de acordo com a diferença. Opcionalmente, considerando separadamente contribuições positivas e negativas, o DeepLIFT também pode revelar dependências que são perdidas por outras abordagens. As pontuações podem ser computadas eficientemente em uma única passagem para trás. O DeepLIFT foi aplicado a modelos treinados em MNIST e em dados genômicos simulados e mostra vantagens significativas sobre os métodos baseados em gradiente [42].

• **Network Dissection**. É um framework para quantificar a interpretabilidade de representações latentes de redes neurais convolucionais (CNNs), avaliando o alinhamento entre unidades ocultas individuais e em conjuntos de conceitos semânticos. Dado qualquer modelo da CNN, o método proposto utiliza um amplo conjunto de dados de conceitos visuais para pontuar a semântica de unidades ocultas em cada camada convolucional intermediária. As unidades com semântica recebem rótulos em vários objetos, partes, cenas, texturas, materiais e cores. O método proposto foi usado para testar a hipótese de que a interpretabilidade das unidades é equivalente a combinações lineares aleatórias de unidades; em seguida, o método foi aplicado para comparar as representações latentes de várias redes quando treinadas para resolver diferentes tarefas de treinamento supervisionado e autossupervionado. Foi analisado o efeito de iterações de treinamento, comparando redes treinadas com diferentes inicializações. Em seguida, foi examinado o impacto da profundidade e largura da rede e medido o efeito do abandono e da normalização do lote da interpretabilidade de representações visuais profundas [43].

• **SUMMIT**. É um sistema interativo que dimensiona e resume e visualiza sistematicamente quais características (features) um modelo de aprendizado profundo aprendeu e como essas características interagem para fazer previsões. A visualização no SUMMIT é executada em navegadores modernos e é de código aberto. A figura 10, apresenta uma visão geral da técnica SUMMIT. Os autores sugerem que essa abordagem de sumarização que constrói representações de classe inteiras é um passo importante para o desenvolvimento de explicações de alto nível para redes neurais [5].

• **YASENN (Yet Another System Explaining Neural Networks)**. É um método de interpretação específico do modelo, que usa uma nova abordagem para a interpretação de redes neurais feed-forward com base no particionamento do espaço de sequências de ativações de neurônios. Esse método tem a capacidade de se concentrar na região de entrada específica e de expressar uma explicação em termos de características diferentes daquelas observadas por uma rede neural.

Tecnicamente, o YASENN destila a rede com um conjunto de árvores de decisão que

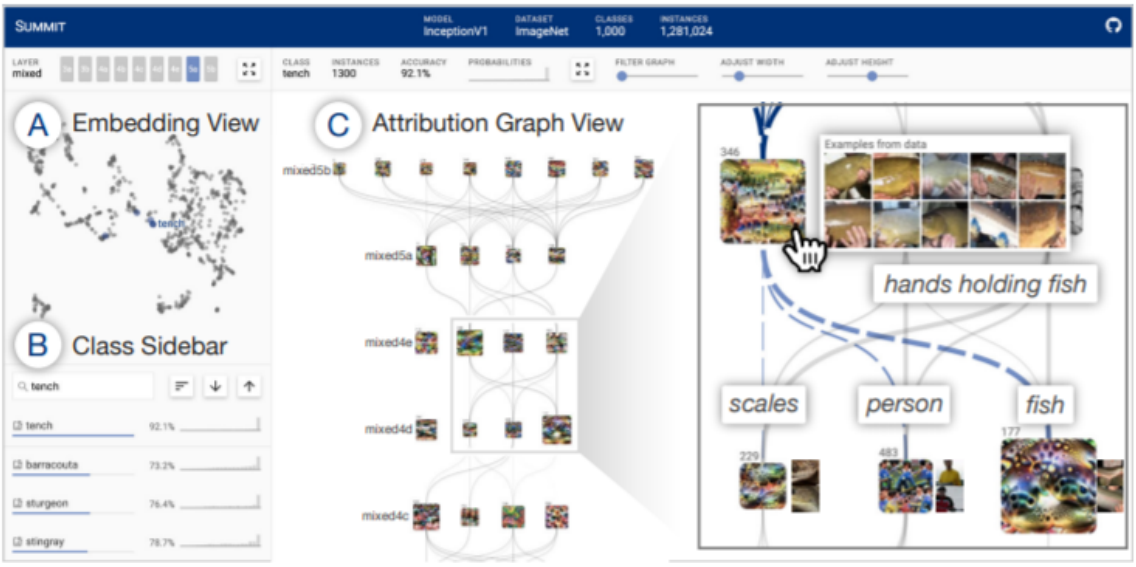

Figura 10 – SUMMIT. Figura retirada de [5].

aumentam o gradiente em camadas e codifica as sequências de ativações de neurônios com índices foliares. O número finito de códigos exclusivos induz a um particionamento do espaço de entrada. Cada partição pode ser descrita de várias maneiras, incluindo o exame de um modelo interpretável (por exemplo, uma regressão logística ou uma árvore de decisão) treinado para discriminar entre os objetos dessas partições [44].

3.3.3 Métodos independentes de modelo (Model-Agnostic Methods)

Métodos independentes de modelo (Model-Agnostic Methods) são métodos que podem ser usados em qualquer modelo de aprendizado de máquina e são aplicados após o treinamento do modelo (post hoc). Os métodos independentes de modelo (agnósticos) geralmente funcionam analisando pares de entrada e saída de características. Esses métodos não podem ter acesso ao modelo interno, como pesos ou informações estruturais. Em teoria, um método independente de modelo pode explicar qualquer tipo de modelo de aprendizagem de máquina, como uma rede neural ou uma simples árvore de decisão.

Aspectos desejáveis de um sistema de explicação independente de modelo são [6]:

Flexibilidade de modelo: Para a maioria das aplicações do mundo real, é necessário treinar modelos precisos e específicos para uma determinada tarefa. Nos métodos independentes de modelo (modelos agnósticos), o modelo é tratado como uma caixa-preta. Flexibilidade do modelo significa que o método de explicação pode funcionar com qualquer modelo de aprendizado de máquina, como florestas aleatórias ou redes neurais profundas.

Flexibilidade da explicação: Se refere à capacidade do método de explicação de usar diferentes representações das características do modelo que está sendo explicado. Por exemplo, em alguns casos, pode ser útil ter uma fórmula linear, em outros casos um gráfico com a importância das características.

Usuários diferentes também podem lidar com diferentes tipos de explicações; um usuário treinado em estatística pode ser capaz de entender uma rede bayesiana, enquanto um modelo linear é mais intuitivo para um usuário leigo. Mesmo que o tipo de explicação seja mantido fixo, os usuários podem tolerar granularidades diferentes em diferentes situações. Por outro lado, mantendo o modelo separado das explicações, é possível adaptar a explicação às necessidades de informações, mantendo o modelo fixo.

Flexibilidade de representação: Se refere à capacidade do método de explicação de usar diferentes representações das características, conforme o modelo que está sendo explicado. Em domínios como imagens, áudio e texto, muitas das características (features) usadas para representar instâncias em soluções do estado da arte não são interpretáveis. As abordagens agnósticas de modelo podem gerar explicações usando características diferentes daqueles usados pelo modelo subjacente.

Menor custo de mudança: A troca de modelos de aprendizagem de máquina é comum em empresas que fazem uso de tais técnicas. Você pode desejar mudar seu modelo baseado em arvores de decisão para uma rede neural com várias camadas por razões técnicas, por exemplo. Quando usamos técnicas de explicações independentes de modelo (agnósticas), mudar o modelo subjacente para um novo é trivial, enquanto o caminho em que as explicações são apresentadas é mantida.

Veja abaixo os principais métodos independentes de modelo (model-agnostic):

• **LIME (Local interpretable model-agnostic explanations)**. É um método post-hoc usado para explicar previsões de qualquer classificador [3]. LIME é uma das bibliotecas mais populares; esse método explica o comportamento local do modelo em torno de algum ponto x. É importante ressaltar que LIME assume que o comportamento do modelo local é muito menos complexo que o global. Se essa suposição for verdadeira, podemos aproximar o comportamento local com um modelo menos complexo e mais interpretável. Esse tipo de modelo é chamado de "substitutos locais", uma vez que eles se aproximam do limite de decisão local de um classificador de caixa-preta. Modelos substitutos locais são modelos interpretáveis usados para explicar previsões individuais de modelos de aprendizado de máquina de caixa-preta.

LIME é um dos representantes mais conhecidos da abordagem de explicação por sim-

plificação, onde os modelos simplificados são apenas representativos de certas seções de um modelo. Explicações locais também são representantes dessa categoria. Quase todas as técnicas que seguem esse caminho para simplificação de modelo são baseadas em técnicas de extração de regras [6].

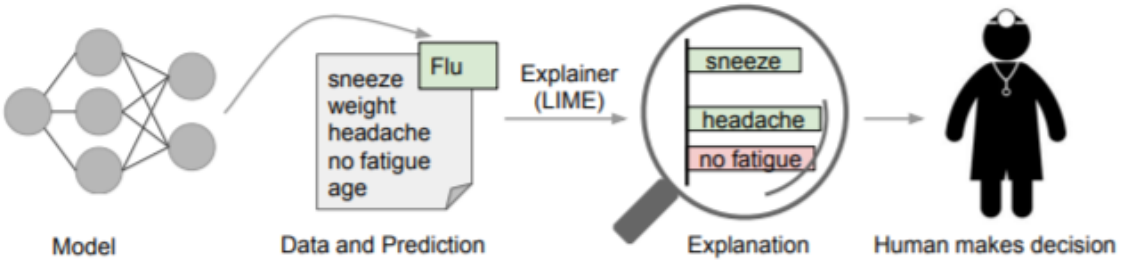

Figura 11 – Previsões individuais no LIME. Figura retirada de [3].

Na figura 11, temos um modelo que prevê se um paciente está com gripe e o LIME destaca os sintomas no histórico do paciente que levaram à previsão. Espirro e dor de cabeça são retratados como contribuindo para a previsão da "gripe", enquanto "sem fadiga"é uma evidência contra ela. Com estes, um médico pode tomar uma decisão informada sobre se deve confiar na previsão do modelo.

No LIME, a interpretabilidade é quantificada com a complexidade das explicações $\{\Omega(g)\}$, onde a medida da complexidade $\{\Omega\}$ pode ser a profundidade da árvore para árvores de decisão ou o número de pesos diferentes de zero para modelos lineares. O modelo $\{f\}$ que está sendo explicado deve retornar valores numéricos $\{f : Rd \rightarrow R\}$, por exemplo, pontuações de probabilidade na classificação [3].

A localidade é definida usando uma medida de proximidade $\{\Omega\}$ entre a instância explicada $\{x\}$ e os pontos perturbados $\{z\}$ em sua vizinhança. A fidelidade local $\{L(f,g,\Omega)\}$ é uma medida de quão infiel o modelo de explicação $\{g\}$ é na aproximação do modelo de previsão $\{f\}$ na localidade definida por $\{\Omega(x,z)\}$. A explicação escolhida minimiza a soma da infidelidade local $\{L\}$ e da complexidade $\{\ \Omega\}$:

$$\text{explanation}(x) = \arg\min_{g \in G} L(f,g,\Pi_x) + \Omega(g) \tag{3.1}$$

A abordagem usa amostragem em torno da instância de explicação $\{x\}$ para desenhar amostras $\{z\}$ ponderado pela distância $\{\pi(x,z)\}$. As amostras formam um conjunto de treinamento para um modelo $\{g\}$ de uma classe de modelo interpretável, por exemplo, um modelo linear. Devido à localidade imposta por π, é esperado que o modelo $\{g\}$ seja uma aproximação fiel de $\{f\}$. LIME usam modelos lineares como uma classe de modelos

interpretáveis {G}, a perda ao quadrado como uma medida de infidelidade local, o número de pesos diferentes de zero como uma medida de complexidade { Ω} e a escolha de pontos de amostra na vizinhança da instância de explicação {x}, de acordo com a distribuição gaussiana da distância entre {x} e o ponto amostrado {z}.

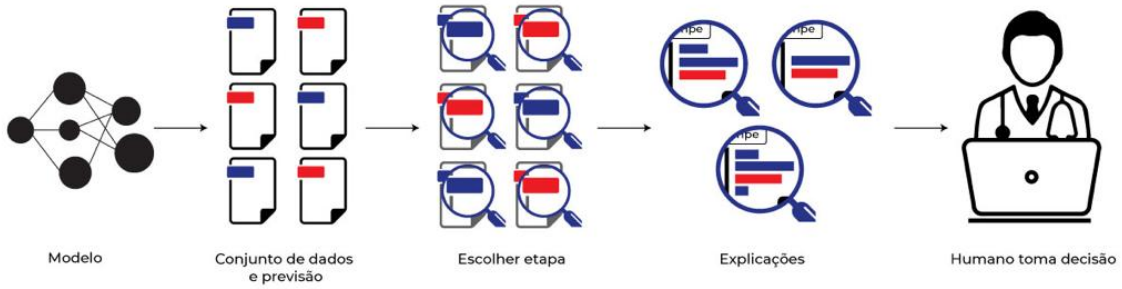

Figura 12 – Explicando um modelo para um tomador de decisão humano. Figura retirada de [6].

O LIME pode trabalhar com dados de texto, imagem ou dados tabulares, que é o foco deste trabalho. Dados tabulares são dados fornecidos em tabelas, com cada linha representando uma instância e cada coluna um característica. No caso de dados tabulares, o LIME cria novas amostras, perturbando cada característica (feature) individualmente, obtendo uma distribuição normal com média e desvio padrão retirado da característica [45].

As amostras do LIME não são coletadas em torno da instância de interesse, mas do centro dos dados de treinamento, o que pode ser problemático. Isso aumenta a probabilidade de o resultado de algumas das previsões de pontos de amostra diferir do ponto de interesse dos dados e de que o LIME possa aprender pelo menos alguma explicação [46].

A figura 13 mostra um exemplo de uma explicação com o LIME de um sistema de aprendizagem de máquina baseado em randon forest que prevê se um time de futebol terá o vencedor do "Homem do jogo"com base nas estatísticas de uma partida da copa do mundo de futebol FIFA 2018. O prêmio "Homem do jogo"é concedido ao melhor jogador do jogo. Os dados foram obtidos do repositório Kaggle [1].

Como interpretar o resultado do LIME? Os valores em laranja são as características mais importantes e as características em azul são os menos importantes na previsão. No nosso exemplo, a característica mais importante foi gols marcados e chutes no alvo.

• **SHAP (SHapley Additive exPlanations)**. É uma estrutura unificada para interpretar previsões, que atribui a cada característica (feature) um valor de importância para uma

[1] https://www.kaggle.com/mathan/fifa-2018-match-statistics

Figura 13 – Técnica LIME. Explicações visuais de um modelo de futebol.

previsão específica. O SHAP [25] unifica vários métodos de interpretabilidade sob o conceito de valores de Shapley [26].

Um valor de Shapley é um conceito da teoria dos jogos, que nos dizem como distribuir de maneira justa o "pagamento"entre as características (features) de um modelo. Uma previsão pode ser explicada assumindo que cada valor de uma característica da instância seja um "jogador"em um jogo em que a previsão é o "pagamento".

Se pensarmos em cada característica (feature) como jogador, a amostra como sua coalizão e a função de pontuação como o valor dessa coalizão, o valor de Shapley para cada característica mostrará como essa característica específica (jogador) afeta a pontuação final. Para avaliar esse valor para uma amostra específica, precisamos avaliar todas as coalizões possíveis de valores de características com e sem essa característica.

Uma inovação que o SHAP traz é que a explicação do valor de Shapley é representada como um método de atribuição de característica aditivo, um modelo linear. O KernelSHAP estima para uma instância x as contribuições de cada valor da característica para a previsão.

Outra variantes para o KernelSHAP é o TreeSHAP. O TreeSHAP é uma variante do SHAP para modelos de aprendizado de máquina baseados em árvores, como árvores de decisão, florestas aleatórias (random forests) e árvores com aumento de gradiente (gradient boosted trees). O TreeSHAP [27] foi introduzido como uma alternativa rápida e específica de modelo ao KernelSHAP [25].

A figura 14 mostra um exemplo de uma explicação com o SHAP (usando o TreeSHAP) de um sistema de aprendizagem de máquina baseado em randon forest que prevê se um time de futebol terá o vencedor do "Homem do jogo"com base nas estatísticas de

Figura 14 – Técnica SHAP. Explicações visuais de um modelo de futebol.

uma partida da copa do mundo de futebol FIFA 2018. O prêmio "Homem do jogo"é concedido ao melhor jogador do jogo.

Como interpretar o resultado do SHAP? Os valores em vermelho são as características mais importantes e as características em azul são os menos importantes na previsão. No nosso exemplo, a característica mais importante foi gols marcados e chutes no alvo.

• **Permutation Importance**. É um método que avalia o impacto de uma característica (feature) no desempenho de um modelo de caixa-preta. O algoritmo mede a importância da característica (feature) observando o quanto diminui a pontuação (métricas de precisão, F1, Recall, etc) quando uma característica não está disponível.

Embora a maneira mais simples de avaliar o impacto da característica na perda alvo seja removê-la do conjunto de dados original e reavaliar a perda, isso exige o retreinamento para quase todos os modelos. Sendo assim, essa abordagem é considerada não prática, porque o treinamento do modelo pode levar muito tempo. Em vez disso, o algoritmo substitui os valores dessa característica por um "ruído aleatório", para que ele perca a relação com o valor alvo.

Para evitar treinar novamente o estimador, o algoritmo poderia remover uma característica apenas da parte de teste do conjunto de dados e calcular a pontuação sem usar essa característica. Portanto, em vez de remover uma característica, o algoritmo o substitui por 'ruído' aleatório - a coluna da característica ainda está lá, mas não contém mais informações úteis. Este método funciona se o ruído for extraído da mesma distribuição que os valores da característica original. A maneira mais simples de obter esse ruído é embaralhar os valores de uma característica, ou seja, usar os valores de característica de outros exemplos - é assim que a importância da permutação é calculada. Veja o algoritmo do Permutation Importance abaixo [47]:

- Inputs: modelo preditivo ajustado m, conjunto de dados tabulares (treinamento ou validação) D.
- Calcule as pontuações de referência s do modelo m nos dados D (por exemplo, a precisão de um classificador ou do R^2 para um regressor).

- Para cada característica j (coluna de D):
  - Para cada repetição K em $1, \ldots, k$:
  - -> Aleatoriamente, embaralhe a coluna j do conjunto de dados D para gerar uma versão corrompida dos dados denominada $\tilde{D}_{k,j}$.
  - -> Calcule a pontuação s_{kj} do modelo m em dados corrompidos $\tilde{D}_{k,j}$.
  - Calcule a importância i_j para característica f_j definida por:

$$i_j = s - \frac{1}{K} \sum_{k=1}^{k} s_{kj} \tag{3.2}$$

A figura 15 mostra um exemplo de uma explicação com o Permutation Importance. Usamos um sistema de aprendizagem de máquina baseado em randon forest que prevê se um time de futebol terá o vencedor do "Homem do jogo"com base nas estatísticas do time. O prêmio "Homem do jogo"é concedido ao melhor jogador do jogo.

Como interpreta o resultado do Permutation Importance? Os valores em direção ao topo são as características mais importantes e as características em direção ao fundo são os menos importantes. No nosso exemplo, a característica mais importante foi gols marcados.

Weight	Feature
0.0750 ± 0.1159	Gols Marcados
0.0625 ± 0.0791	Escanteio
0.0437 ± 0.0500	Distancia Percorrida (Kms)
0.0375 ± 0.0729	No alvo
0.0375 ± 0.0468	Tiro Livre
0.0187 ± 0.0306	Bloqueado
0.0125 ± 0.0750	Passes Certos %
0.0125 ± 0.0500	Amarelo
0.0063 ± 0.0468	Saves
0.0063 ± 0.0250	Impedimentos
0.0063 ± 0.1741	Fora do Alvo
0.0000 ± 0.1046	Passes
0 ± 0.0000	Vermelho
0 ± 0.0000	Amarelo & Vermelho
0 ± 0.0000	Gols in PSO
-0.0312 ± 0.0884	Faltas Cometidas
-0.0375 ± 0.0919	Tentativas
-0.0500 ± 0.0500	Posse Bola %

Figura 15

O método é mais adequado para calcular a importância de características quando não temos um grande quantidade de características (features), caso contrário, pode ser muito custoso computacionalmente [25–27, 47].

• **Anchors**. É um método independente de modelo proposto pelos criadores do LIME [3] que explica o comportamento de modelos complexos com regras de alta precisão,

que representam condições locais "suficientes"para previsões. O algoritmo do Anchors gera explicações chamadas "âncoras". Essas âncoras fazem parte da amostra explicada, ancorando-a em um rótulo específico produzido pelo modelo para esta amostra explicada. Basicamente, cada âncora é um conjunto de predicados chamado A presente na amostra explicada. Esses predicados podem estar em qualquer forma e geralmente dependem do tipo de dados que está sendo explicado. O conjunto de predicados é chamado de âncora se ancora a amostra em um rótulo. Em outras palavras, se alterarmos a amostra de várias maneiras diferentes, de acordo com alguma distribuição de perturbação, que chamaremos de D, sem quebrar os predicados, e o rótulo não mudar, teremos encontrado a nossa âncora [48].

• **Gráfico de dependência parcial (partial dependence plot (PDP))**. O gráfico de dependência parcial é uma ferramenta que ajuda visualizar o efeito marginal que uma ou duas características têm sobre a previsão de um modelo de aprendizado de máquina. Um gráfico de dependência parcial pode mostrar se a relação entre o alvo e uma característica é linear, monotônica ou mais complexa [49].

• **ALE plots (Accumulated local effects)**. É uma ferramenta para visualizar os efeitos de variáveis preditoras e seus efeitos de interação de ordem inferior em modelos de aprendizado supervisionado. ALE plots são computacionalmente mais baratos que outros métodos com abordagem similares [50].

• **ICE (Individual Conditional Expectation)**. É uma ferramenta para visualizar o modelo estimado por qualquer algoritmo de aprendizagem de máquina supervisionado. Os gráficos da ferramenta ICE exibem uma linha por instância que mostra como a previsão da instância muda quando uma característica é alterada. Os gráficos de ICE refinam o gráfico de dependência parcial (PDP), representando graficamente a relação funcional entre a resposta prevista e a característica para observações individuais. Os gráficos de ICE destacam a variação nos valores ajustados ao longo da faixa de uma co-variável, sugerindo onde e até que ponto heterogeneidades podem existir. A ferramenta também fornece um conjunto de plotagem para análise exploratória de dados [51].

• **LIVE (Local Interpretable Visual Explanations)**. É uma implementação alternativa do método LIME para problemas de regressão. Um dos principais objetivos do LIVE é fornecer ferramentas para visualização, para o entendimento de modelos complexos. O método de exploração local e o manuseio de entradas interpretáveis são alterados, assim como o LIME. O conjunto de dados para exploração local é simulado perturbando a instância explicada, uma característica por vez. As Variáveis originais são usadas como entradas interpretáveis. A interpretabilidade da explicação local vem de um relaciona-

mento tratável entre entradas e a resposta prevista [52].

• **BreakDown**. É um pacote, cujo objetivo principal é decompor as previsões do modelo em partes que podem ser atribuídas a variáveis específicas. O BreakDown atribui as características com base em respostas condicionais de um modelo de caixa-preta. É simples para modelos lineares e para modelos mais gerais (aditivos) [52].

• **MCR (Model Class Reliance)**. É uma técnica baseada em medida de importância variável para qualquer classe de modelo de aprendizado de máquina [53].

• **DALEX**. É uma coleção consistente de explicadores para modelos preditivos e caixas-pretas. Cada explicador é uma técnica para explorar um modelo de caixa preta. Esses explicadores são implementados no pacote DALEX para linguagem estatística R [54].

•**Instancewise feature**. É uma metodologia para interpretação de modelos. O método é baseado no aprendizado de uma função para extrair um subconjunto de características que são mais informativos para cada exemplo. Esse seletor de características é treinado para maximizar as informações mútuas entre as características selecionados e a variável de resposta, onde a distribuição condicional da variável de resposta dada a entrada é o modelo a ser explicado [55].

3.3.4 Métodos baseados em exemplos (Example-Based Explanations)

Explicações baseadas em exemplos são métodos que selecionam instâncias do conjunto de dados para explicar o comportamento dos modelos de aprendizado de máquina ou para explicar a distribuição de dados subjacente.

De acordo com Molnar [18], as explicações baseadas em exemplos são na maioria dos casos independentes de modelo, porque tornam qualquer modelo de aprendizado de máquina mais interpretável. A diferença para os métodos independentes de modelo é que os métodos baseados em exemplo explicam um modelo selecionando instâncias do conjunto de dados e não criando resumos das características.

Os principais tipos de métodos de explicações baseadas em exemplos são: Explicações contrafatuais (Counterfactual explanations), Instâncias contraditórios (Adversarial examples), Protótipos e críticas (Prototypes and criticism) e Instâncias influentes (influential examples).

Explicações contrafatuais (Counterfactual explanations)

Explicações contrafatuais nos dizem como uma instância precisa mudar para alte-

rar significativamente sua previsão. Ao criar instâncias contrafatuais, aprendemos sobre como o modelo faz suas previsões e pode explicar previsões individuais [56].

Instâncias contraditória (Adversarial examples)

Uma instância contraditória é uma instância com pequenas perturbações intencionais nas características que fazem com que um modelo de aprendizado de máquina faça uma previsão falsa. As instâncias contraditórias são similares a explicações contrafatuais (Counterfactual Explanations). Instâncias contraditórias são contrafatuais usadas com o objetivo de enganar modelos de aprendizado de máquina, pois a ênfase está em inverter a previsão e não interpretá-la ou explicá-la [57].

Protótipos e críticas (Prototypes and criticism)

O objetivo das críticas é fornecer informações juntamente com os protótipos, especialmente para pontos de dados que os protótipos não representam bem. Protótipos são uma seleção de instâncias representativas dos dados e críticas são instâncias que não são bem representadas por esses protótipos. Protótipos e críticas podem ser usados independentemente de um modelo de aprendizado de máquina para descrever os dados, mas também podem ser usados para criar um modelo interpretável ou para interpretar um modelo de caixa-preta [56].

Instâncias influentes (influential examples)

Instâncias influentes são os pontos de dados de treinamento que foram os mais influentes para os parâmetros de um modelo de previsão ou para as próprias previsões. Uma instância de treinamento é "influente", quando sua exclusão dos dados de treinamento altera consideravelmente os parâmetros ou previsões do modelo. Identificar e analisar instâncias influentes possibilita "depurar"modelos de aprendizado de máquina. Assim é possível encontrar problemas com os dados e entender melhor o comportamento e previsões do modelo [58].

4. A perspectiva do usuário sobre as técnicas de Inteligência Artificial Explicável

Com o objetivo de reforçar a relevância do tema abordado nesta pesquisa, passa-se a apresentar informações sobre a avaliação da explicabilidade sob a perspectiva do usuário, bem como apresentar referências a outros trabalhos da literatura sobre a explicabilidade e a avaliação das técnicas de inteligência artificial explicável, senão vejamos:

4.1 Avaliação humana das técnicas de IA explicável na perspectiva do usuário especialista do domínio

Atualmente, é possível criar sistema de inteligência artificial bastante preciso com base em aprendizagem de máquina, haja vista a existência de técnicas recentes de aprendizagem de máquina como, por exemplo, os métodos ensembles e as redes neurais profundas de última geração.

Todavia, é relevante dizer que por mais preciso que seja o resultado apresentado por um sistema de inteligência artificial ele de nada servirá se não for utilizado por quem dele possa se beneficiar.

Neste tocante, encontra-se um ponto que merece atenção, qual seja: como fomentar o uso do sistema de inteligência artificial? É empírico constatar que o uso de qualquer sistema requer confiança em sua metodologia, sobretudo quando se trata de um sistema de inteligência artificial capaz de influenciar importantes decisões, principalmente dependendo do âmbito em que for utilizado. Desta forma, naturalmente, acolhe-se aquilo em que se confia e rejeita-se aquilo em que se desconfia. Pois bem.

Isto porque, em muitos cenários realistas, modelos de inteligência artificial tendenciosos podem ser utilizados e trazerem à tona efeitos nefastos como, por exemplo, quando utilizado para traçar perfil de criminosos em potencial ou para definir pontuação de risco

de sentença judicial ou, ainda, para definir pontuação de crédito ou detecção de fraude ou para fazer avaliação de saúde ou para concessão de empréstimo ou, ainda, para definir padrão na condução de carros autônomos e muito mais. Como se pode notar a compreensão e a interpretação dos resultados dos modelos de aprendizagem de máquina são fundamentais em todos os casos acima apontados e muitos outros aqui não elencados [59].

Neste tocante, cumpre registrar que, ao lidar com problemas de aprendizado de máquina, os projetistas geralmente tendem a se fixarem nas métricas de desempenho do modelo como, por exemplo, acurácia, precision, recall e F1-score. Contudo, vale dizer que as métricas de desempenho não atendem aos usuários finais dos sistemas de inteligência artificial, pois dizer que um sistema de IA baseado em aprendizagem de máquina de diagnóstico de câncer de colo de útero tem 0.91 de acurácia é absolutamente insuficiente para o usuário do domínio. Conclui-se, pois, se de suma importância entender o que leva um sistema de aprendizagem de máquina a tomar as decisões por ele apresentadas.

Constata-se, historicamente, que a questão começou a ser estudada a partir do surgimento de técnicas de inteligência artificial explicável independentes de modelo (agnósticas) como são: o LIME e o SHAP, técnicas abordadas detalhadamente no capítulo anterior. Nota-se, ainda, que o esforço maior nas pesquisas realizadas na área de inteligência artificial explicada (XAI) está relacionado à criação de novas técnicas de XAI. Por outro lado, verifica-se que poucas são as pesquisas realizadas com o objetivo de entender a qualidade da explicação trazida por essas técnicas.

Portanto, mais uma razão para se avaliar a relevância humana e a compreensibilidade das explicações das pesquisas realizadas na área de inteligência artificial explicada (XAI).

Neste cenário, importa trazer à baila a exposição de Gilpin [17] que afirma que a avaliação humana em inteligência artificial explicável é o ato de avaliar explicações por razoabilidade, pois é assim que uma explicação corresponde às expectativas humanas. Continua afirmando que a avaliação humana também pode avaliar a integridade ou a integridade da tarefa substituta do ponto de vista de permitir que uma pessoa preveja comportamento do modelo original; ou de acordo com a utilidade em revelar vieses de modelo para uma pessoa.

Do mesmo modo, apresenta-se o pensamento de Gunning e colegas [59] que afirmam que a medição confiável e consistente dos efeitos das explicações ainda é uma questão de pesquisa em aberto.

Já de acordo com Miller [9], a maioria dos trabalhos em inteligência artificial explicável usa apenas a intuição dos pesquisadores/projetistas do que constitui uma explicação

"boa".

Diante deste cenário de muitas suposições e nenhuma métrica de medição das técnicas de inteligência artificial explicável nasce a inspiração desta pesquisa.

Nesta esteira, tem-se como objetivo principal da presente pesquisa avaliar qualitativamente a capacidade das técnicas de inteligência artificial explicável agnósticas para oferecer informações úteis, bem como para esclarecer ao usuário especialista do domínio os resultados das técnicas aplicadas e, com isso, garantir confiança em seus resultados e melhor e maior aplicação das técnicas em diversas áreas.

4.2 Trabalhos relacionados

Como demonstrativo, passa-se a apresentar os principais trabalhos relacionados a esta pesquisa com foco na avaliação dos sistemas XAI decorrente de levantamento bibliográfico de trabalhos recentes. Neste ponto, importa registrar que foram pesquisados apenas artigos publicados a partir de 2016, uma vez que se trata de tema novo e, portanto, ainda pouco estudado.

Inicia-se a apresentação com o trabalho de Lipton [35], que buscou em sua pesquisa refinar o discurso sobre a interpretabilidade sob a ótica das motivações subjacentes ao interesse por ela. Uma das grandes contribuições para o nosso estudo que Lipton apresenta são as motivações subjacentes ao interesse pela interpretabilidade (confiança, causalidade, transferibilidade, informatividade, tomada de decisão justa e ética) que, segundo ele, são diversas e ocasionalmente divergentes. Ao passo que abordar as propriedades e técnicas do modelo que se pensa conferir interpretabilidade, identifica a transparência para os seres humanos (simulabilidade, decomposição e transparência algorítmica) e explicações post-hoc (explicações em texto, visualização, explicações locais e explicações por exemplo) como noções concorrentes.

Por outro lado, segundo Biran e Cotton [10], explicabilidade é algo fortemente relacionado à noção de interpretabilidade: um sistema interpretável seria aquele cujas operações são compreensíveis para nós humanos, seja por meio da inspeção do sistema, seja por meio de alguma explicação produzida durante o seu funcionamento. Além disso, eles estabelecem uma distinção entre interpretabilidade e a noção de justificação, cujo objetivo desta última seria explicar por que a decisão tomada pelo sistema pode ser aceita como uma boa decisão. Concluindo, pois, que justificabilidade e interpretabilidade são capacidades complementares.

Já Doshi-Velez e Been Kim [19] propuseram uma revisão com o objetivo de traçar um caminho em direção à noção de avaliação rigorosa da interpretabilidade e, nesta toada, passaram a considerar cenários em que se considera a interpretabilidade necessária e o porquê de tal importância. Além disso, apresentaram proposta de uma taxonomia para a avaliação da interpretabilidade (application-grounded, human-grounded e functionally grounded), bem como ventilaram importantes questões ainda abertas e propuseram questões específicas para pesquisadores atuantes na área.

Para Herman [23], a interpretabilidade funcional pode estar correlacionada à função cognitiva e às preferências do usuário e se essa correlação realmente existir, a avaliação e a otimização usando métricas funcionais podem traduzir viés cognitivo implícito nas explicações, ameaçando a transparência. Mas, por outro lado, propõe direções de pesquisa em potencial para desambiguar a função cognitiva e os modelos de explicação e, assim, garantir o equilíbrio entre precisão e interpretabilidade.

De acordo com Abdul e colegas [60] fizeram uma análise completa da literatura com o intuito encontrarem tópicos relacionados à XAI e a relação entre os tópicos. Para isso, usaram a visualização de modelo de tópico de palavras-chave e rede de citações para apresentar uma visão holística dos esforços de pesquisa em XAI, incluindo privacidade e justiça, agentes inteligentes, sistemas sensíveis ao contexto, responsabilidade algorítmica, psicologia cognitiva, aprendizado de software, dentre outros.

Noutro giro, Guidotti e colegas [34] fornecem uma classificação dos principais problemas abordados na literatura em relação à noção de explicação e ao tipo de sistema de 'caixa preta', muito embora a literatura relate muitas abordagens destinadas a superar essa fraqueza crucial, muitas vezes com o custo de sacrificar a precisão em detrimento da interpretabilidade. Eles apontam que as aplicações nas quais os sistemas de decisão de 'caixa preta' podem ser usadas são diversas, sendo certo que cada abordagem é, em regra, desenvolvida para fornecer uma solução para um problema específico e, consequentemente, consegue delinear explícita ou implicitamente sua própria definição de interpretabilidade e explicação.

Ainda segundo Guidotti e colegas, dada uma definição de problema, um tipo de 'caixa preta' e uma explicação desejada, a pesquisa deve ajudar o pesquisador a encontrar as propostas mais úteis ao seu próprio trabalho. Portanto, a classificação proposta de abordagens para abrir modelos de 'caixa preta' também deve ser útil para colocar em perspectiva as muitas questões abertas da pesquisa em inteligência artificial explicável, até porque a ausência de explicação constitui uma questão prática e ética.

Para Doran e colegas [61], há 03 (três) tipos de sistemas de inteligência artificial explicável: a) sistemas opacos; b) sistemas interpretáveis e c) sistemas compreensíveis. Os sistemas opacos não oferecem insights sobre seus mecanismos algorítmicos. Já os sistemas interpretáveis são aqueles em que os usuários podem analisar matematicamente seus mecanismos algorítmicos. Por fim, os sistemas compreensíveis são aqueles que emitem símbolos, permitindo explicações orientadas pelo usuário sobre como chegar a uma conclusão, ou seja, compreender por que uma certa saída está associada a uma certa entrada. Deste modo, os autores afirmam que compreensibilidade e interpretabilidade seriam capacidades complementares.

Mohseni e colegas [62] sugerem que há diferentes objetivos de avaliação na pesquisa de aprendizado de máquina interpretável, através de uma revisão minuciosa das metodologias de avaliação usadas na pesquisa de explicação de máquina nos campos da interação humano-computador, análise visual e aprendizado de máquina.

Dosilovic e colegas [63] mostram em sua pesquisa os avanços na interpretabilidade e na explicabilidade do aprendizado de máquina sob o prisma do aprendizado supervisionado. De acordo com os autores, grande parte do trabalho recente está na área de aprendizado profundo, por um lado por conta dos ganhos notáveis de desempenho destes modelos e, por outro lado, por sua opacidade intrínseca. Na sequência, os autores iniciam uma discussão sobre a conexão entre explicabilidade com a inteligência artificial e apresentam propostas para novas direções de pesquisa.

Já Miller [9] sugere em seu estudo uma abordagem diferente, isto porque, a partir de uma revisão da literatura, ele demonstra que há um escopo considerável para introduzir mais resultados das ciências sociais e comportamentais na IA explicável, apresentando alguns resultados importantes destes campos que são relevantes para a IA explicável.

O autor defende que, embora o ressurgimento da IA explicável seja positivo, argumenta que a maioria de nós, a exemplo de pesquisadores da IA, constrói agentes explicativos para si próprios e não para os usuários pretendidos.

Miller segue argumentando que a IA explicável tem maior probabilidade de sucesso se pesquisadores e profissionais entenderem, adotarem, implementarem e melhorarem modelos dos vastos e valiosos corpos de pesquisa em filosofia, psicologia e ciência cognitiva e que a avaliação destes sistemas seja mais voltada para as pessoas do que para a tecnologia.

Lage e colegas [64] avançam no sentido de fornecerem uma base empírica para quais tipos de explicações os seres humanos podem utilizar. Nesta linha, concentraram-se em

conjuntos de decisões e, com isso, determinaram como 03 (três) tipos diferentes de complexidade das explicações, quais sejam: 1- duração das cláusulas e explicações; 2- número e apresentação de blocos cognitivos (conceitos recém-definidos) e 3- repetições variáveis, afetam a capacidade dos humanos de usarem essas explicações em 03 (três) tarefas diferentes, 02 (dois) domínios distintos e 03 (três) métricas de desempenho diversas.

Concluem que o tipo de complexidade é importante, pois os blocos cognitivos afetam mais o desempenho do que as repetições variáveis e essas tendências são consistentes entre tarefas e domínios, sugerindo que possa existir alguns princípios comuns de design para sistemas de explicação.

Mittelstadt e colegas [14] focaram na distinção entre trabalhos recentes sobre interpretabilidade no aprendizado de máquina e explicações em filosofia e sociologia. Segundo eles, esse sistema pode ser entendido como um "kit faça você mesmo"para explicações, permitindo que o profissional responda diretamente "perguntas do tipo se"ou gere explicações contrastantes sem assistência externa. Embora se trate de um sistema com habilidade valiosa, tem-se que fornecer tal sistema como explicação parece mais difícil do que o necessário e, além disso, outras formas de explicação podem não ter as mesmas vantagens e desvantagens. Os autores comparam as diferentes escolas de pensamento sobre o que faz uma explicação e sugerem que o aprendizado de máquina possa se beneficiar da visualização do problema.

Gilpin e colegas [17] descrevem conceitos fundamentais de explicabilidade e mostram como eles podem ser usados para classificar a literatura existente em um esforço para criar boas práticas e identificar desafios abertos.

Conforme os autores, embora essas explicações sejam importantes para garantir a equidade do algoritmo, é preciso identificar possíveis lacunas e/ou problemas nos dados de treinamento a fim de garantir que os algoritmos funcionem conforme o esperado e as explicações produzidas por esses sistemas não sejam padronizadas e nem sistematicamente avaliadas. Eles apontam que as abordagens atuais dos métodos explicativos, especialmente para redes neurais profundas, são insuficientes. Por fim, concluem o estudo com sugestões de orientações futuras para a inteligência artificial explicável.

Gunning e colegas [59] apresentam em artigo conceitos fundamentais de inteligência artificial explicável e argumentam que a explicabilidade é essencial para que os usuários compreendam, confiem e gerenciem efetivamente as aplicações de inteligência artificial.

Os autores indicam um importante desafio em inteligência artificial explicável, qual seja, a avaliação e a medição da explicabilidade. Em que pese várias maneiras de avaliar

e medir a eficácia de uma explicação tenham sido propostas nos últimos anos, atualmente não há meios comuns para medir se um sistema XAI é mais inteligível ou não para um usuário do que um sistema não XAI.

Algumas dessas avaliações são medidas subjetivas do ponto de vista do usuário como, a exemplo da satisfação do usuário, que pode ser aferida por meio de uma classificação subjetiva da clareza e utilidade de uma explicação.

Portanto, medidas mais objetivas para a eficácia de uma explicação podem ser o desempenho da tarefa; ou seja, se a explicação melhora a tomada de decisão do usuário. De acordo com os autores a medição confiável e consistente dos efeitos das explicações ainda é uma questão de pesquisa em aberto e, por isso, apontam algumas questões e desafios na área de inteligência artificial explicável, senão vejamos:

a) As explicações devem partir de computadores versus ser orientado às pessoas;

b) Precisão versus interpretabilidade, ou seja, a busca de um equilíbrio entre precisão e interpretabilidade;

c) Usar abstrações para simplificar explicações;

d) Explicar competências versus explicar decisões, ou seja, a necessidade de ajudar o usuário final a entender as competências dos sistemas de IA em termos de quais competências um sistema específico de IA possui, como as competências devem ser medidas e se um sistema de IA tem pontos cegos.

Paez [65] diz que o objetivo de fornecer uma explicação de um sistema de aprendizagem de máquina ou de uma decisão é torná-lo compreensível para seus stakeholders e que sem uma compreensão prévia do que significa dizer que um agente entende um sistema ou uma decisão, as estratégias explicativas não terão um objetivo bem definido.

De acordo com o autor, fornecer um objetivo mais claro para a XAI e o foco no entendimento também permite relaxar a condição de exequível da explicação, impossibilitando o cumprimento de muitos sistemas de aprendizado de máquina, devendo haver concentração nas condições pragmáticas que determinam o melhor ajuste entre um sistema e os métodos e dispositivos implantados para melhor entendê-lo. Argumenta, ainda, que a busca por sistemas explicáveis e decisões interpretáveis em IA deve ser reformulada em termos do projeto mais amplo de oferecer uma explicação pragmática e naturalista do entendimento em IA.

Paez chega a essa conclusão após exame dos diferentes tipos de entendimento discu-

tidos na literatura filosófica e psicológica, onde conclui que os sistemas interpretativos ou de aproximação não apenas fornecem a melhor maneira de alcançar a compreensão objetiva de um sistema de aprendizado de máquina, mas também são uma condição necessária para alcançar a interpretabilidade post hoc. A conclusão de Paez é parcialmente baseada nas deficiências da abordagem puramente funcionalista da interpretabilidade post hoc e que parece ser predominante na literatura mais recente.

Barredo Arrieta e colegas [7] realizaram uma revisão sistemática da literatura em torno da Inteligência Artificial explicável (XAI). No estudo, primeiramente, esclareceram os diferentes conceitos subjacentes à explicabilidade, bem como os diversos propósitos que motivam a busca por métodos de aprendizagem de máquina mais interpretáveis, sendo pois:

1) algoritmos de aprendizagem de máquina que apresentam algum grau de transparência e, portanto, podem ser interpretados por eles mesmos; e

2) técnicas XAI post-hoc criadas para tornar os modelos de aprendizagem de máquina mais interpretáveis.

Esta análise da literatura produziu uma taxonomia global de diferentes propostas relatadas pela comunidade, classificando-as de acordo com critérios uniformes.

Dada a prevalência de contribuições relacionadas à explicabilidade dos algoritmos de Aprendizado Profundo foi examinado com afinco a literatura que trata dessa família de algoritmos, dando origem a uma taxonomia alternativa que se conecta mais estreitamente aos domínios específicos nos quais a explicabilidade pode ser realizada para os algoritmos de aprendizagem profunda.

Assim, os autores discutiram o conceito da IA responsável, paradigma que impõe uma série de princípios de IA a serem observados quando da implementação dos modelos de IA na prática, incluindo, mas não se limitando, à justiça, à transparência e à privacidade. Ademais, também foi discutido as implicações da adoção de técnicas XAI no contexto da fusão de dados, revelando potencial da XAI de comprometer a privacidade dos dados protegidos envolvidos no processo de fusão. Da mesma forma, as implicações da XAI no âmbito da justiça também foram objeto de discussão detalhada, além de reflexões sobre o futuro da XAI, além de um entendimento adequado das potencialidade e questões abertas pelas técnicas XAI.

A visão dos autores é de que a interpretabilidade dos sistemas de aprendizagem de máquina deve ser tratada em conjunto com os requisitos e restrições relacionados à pri-

vacidade dos dados, à confidencialidade, à justiça e à responsabilidade, ou seja, a implementação e o uso responsável dos métodos de IA em organizações e instituições em todo o mundo só será garantida se todos os princípios de IA forem estudados em conjunto.

Todavia, ressalto que há uma lacuna de estudos em inteligência artificial explicável que visa uma avaliação humana de técnicas agnósticas por usuários especialista no domínio, ou seja, caracterizar se as técnicas agnósticas oferecem explicabilidade. Dentre os trabalhos encontrados na literatura, a pesquisa de Mohseni e Ragan [66], Weerts e colegas [67] e Wang e outros colegas [68] foi a que mais se aproximou deste fim.

Isto porque, Mohseni e Ragan [66] propuseram uma nova metodologia de avaliação para explicações detalhadas de classificadores de texto e imagem. O propósito deles era avaliar a relevância e a adequação de explicações locais sobre os resultados do aprendizado de máquina. Os metadados de explicação neste benchmark são gerados a partir de anotações do usuário, de amostras de imagem e texto. Eles descrevem o benchmark e demonstram sua utilidade por uma avaliação quantitativa das explicações geradas por um recente algoritmo de aprendizado de máquina. Esta pesquisa demonstra como a avaliação baseada em humanos pode ser usada como uma medida para qualificar explicações locais de aprendizado de máquina.

Enquanto Weerts e colegas [67] apresentam os resultados de uma avaliação humana da técnica de IA explicável SHAP. Eles autores realizaram um experimento com 03(três) grupos diferentes de participantes (sendo 159 participantes, no total) que possuíam conhecimentos básicos de inteligência artificial explicável.

Eles realizaram uma análise qualitativa das reflexões registradas dos participantes do experimento, realizando processamento de alerta com e sem informações da técnica de inteligência artificial explicável, o SHAP. Os autores testaram estatisticamente se havia diferença significativa entre as métricas do utilitário de tarefas, entre as tarefas para as quais uma explicação estava disponível e entre tarefas nas quais ela não foi fornecida.

Ao contrário de intuições comuns, não foi encontrado uma diferença significativa no desempenho da tarefa de processamento de alertas quando uma explicação provida pelo método SHAP está disponível e quando o método SHAP não está disponível.

Wang e colegas [68] propuseram um framework conceitual orientado pela teoria para desenvolver IA (XAI) centrada no usuário. Os autores utilizaram filosofia, psicologia cognitiva e inteligência artificial. O objetivo do framework é ajudar os desenvolvedores a criar sistemas baseados em IA explicáveis e centrados no usuário, detalhando os recursos técnicos do XAI e os conectando aos requisitos do raciocínio humano.

De acordo com os autores, ao usar o framework, é possível identificar caminhos de como explicações específicas podem ser úteis, como certos métodos de raciocínio falham porque ligados a um viés cognitivo e, por fim, como aplicar diferentes elementos do XAI para mitigação dessas falhas.

É bem verdade que o framework tenha sido criado especificamente para o domínio médico, mas os autores fizeram uma série de recomendações sobre como refinar ainda mais os projetos XAI a fim de melhorar a interpretabilidade humana e, com isso, permitir que essas recomendações possam ser aplicadas a outros domínios.

Por fim, conclui-se que nenhum dos trabalhos acima expostos de maneira resumida realizou uma avaliação humana qualitativa com o objetivo de comparar a explicabilidade e a interpretabilidade de técnicas populares de inteligência artificial explicável agnósticas, como o SHAP, LIME e Permutation Importance, sobretudo entre usuários especialistas em um domínio crítico, sendo, pois, este o diferencial da presente pesquisa.

5. Metodologia de Pesquisa

A presente pesquisa objetiva apresentar as técnicas de inteligência artificial explicável sob o ponto de vista do usuário especialista no domínio e, por isso, adotou-se uma abordagem qualitativa, pois se pretende compreender o comportamento do especialista quando ele recebe um resultado e uma explicação, ambas advindas de um sistema inteligente.

Métodos de pesquisa qualitativos são estratégias de investigação empírica que investigam fenômenos dentro de um contexto de vida real [69] e, portanto, mais adequados ao estudo do entendimento das técnicas de inteligência artificial explicável no domínio médico (oncologia).

A pesquisa possui caráter exploratório, uma vez que foram feitos estudos com o objetivo de obter uma maior compreensão e entendimento da explicabilidade de técnicas de inteligência artificial explicável, bem como detectar eventuais dificuldades em seu uso e verificar oportunidades de melhoria por especialistas no domínio médico.

Cumpre dizer que a pesquisa utilizou o Método de Explicitação do Discurso Subjacente (MEDS) [70] - que é um método qualitativo -, já que a coleta de dados foi obtida por meio de entrevistas presenciais, realizadas em um modelo típico de conversas cotidianas extraídas de contextos informais.

Registra-se que a abordagem da pesquisa adotada foi a teoria fundamentada nos dados, constituído pela interpretação e análise das entrevistas com os especialistas no domínio médico (oncologia). Frisa-se que a Teoria Fundamentada em Dados (Grounded Theory) é uma metodologia sistemática para pesquisa qualitativa que objetiva gerar teorias com base nos dados obtidos ao longo de uma pesquisa.

Portanto, quando se gera uma teoria que explica os dados coletados significa dizer que ela gera uma "teoria fundamentada nos dados" (grounded theory).

O objetivo desta pesquisa é avaliar qualitativamente o entendimento quanto ao resul-

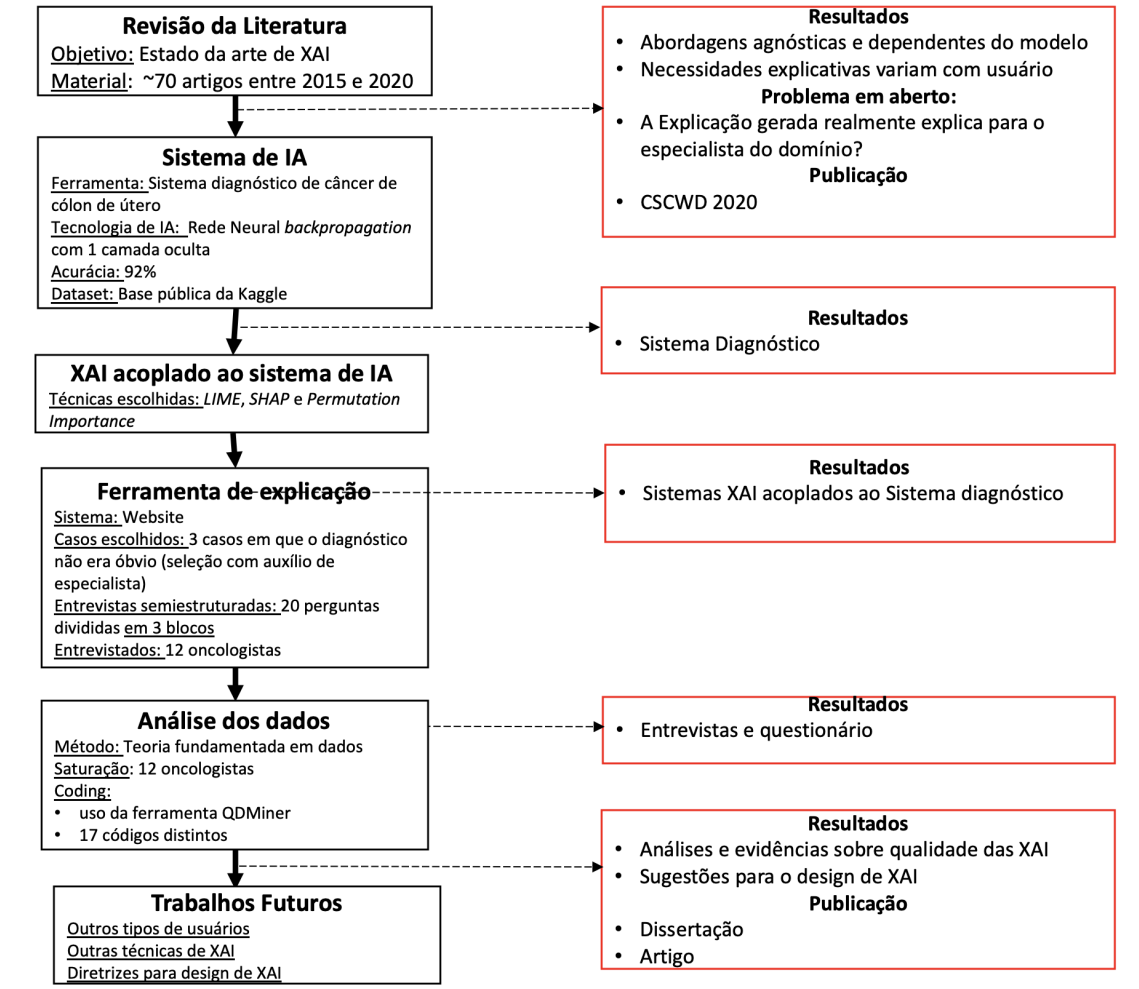

Figura 16 – Metodologia de Pesquisa

tado da explicação de 03 (três) técnicas de inteligência artificial explicável, sendo certo que a pesquisa foi desenvolvida em 07 (sete) etapas, são elas:

1. Revisão da literatura;

2. Definição do domínio, conjunto de dados relacionado ao domínio e construção do modelo de aprendizagem de máquina;

3. Escolhas das técnicas de inteligência artificial explicável;

4. Definição do questionário e do método de coleta de dados;

5. Realização das entrevistas;

6. Análise dos resultados obtidos nas entrevistas;

7. Elaboração das recomendações para o desenvolvimento de sistemas de IA explicável e sugestões para implantação no domínio médico.

5.1 Etapa de revisão da literatura

Nesta primeira etapa, realizou-se uma revisão da literatura para encontrar trabalhos relacionados à inteligência artificial explicável, ao aprendizado de máquina interpretável e a conceitos relacionados ao escopo deste trabalho.

Com esta revisão, foi criada uma string de consulta sem um protocolo específico, já que não foi encontrado um protocolo específico para a área de pesquisa. As fontes do banco de dados foram ACM Digital Library, IEEE Xplorer e Scopus. O primeiro e o segundo bancos de dados foram selecionados para fornecer ampla cobertura no campo Ciência da Computação e o Scopus foi o banco escolhido por ser um dos maiores bancos de dados de resumos e citações da literatura revisada por pares no mundo.

Selecionou-se aproximadamente 70 (setenta) artigos e trabalhos atuais das principais conferências e periódicos de ciência da computação nos campos da inteligência artificial e aprendizado de máquina, entre 2014 e 2019, diretamente relacionados às palavras-chave: "inteligência artificial explicável", "XAI", "aprendizado de máquina interpretável", "interpretabilidade", "explicabilidade", "imparcialidade", "responsabilidade", "transparência".

Na seleção de artigos, estabeleceu-se filtros diretamente nos formulários de pesquisa dos bancos de dados, como pesquisa apenas nos campos de título, resumo e palavra-chave. Além disso, utilizou-se filtros nos artigos restritos à área de Ciência da Computação e, quando possível, filtrou-se apenas artigos de periódicos, conferências (artigos completos) e capítulos de livros diretamente no formulário de extração do banco de dados.

5.2 Etapa de definição do domínio, seleção do conjunto de dados e o método de aprendizagem de máquina

Esta seção apresenta o domínio, conjunto de dados relacionado ao domínio e o modelo de aprendizagem de máquina.

5.2.1 Definição do domínio

Na definição do domínio, o objetivo era a utilização de um domínio crítico, onde a interpretabilidade e a explicabilidade dos resultados são considerados essenciais para a confiança dos resultados.

Neste ponto, relaciona-se a seguir os critérios que embasaram a escolha do domínio

médico:

- Disponibilidade de dados abertos em repositório de ciência de dados;

-Importância do domínio (oncologia);

- Crescente introdução da inteligência artificial na área médica;

- Domínio que exige confiabilidade.

Oncologia

A oncologia ou cancerologia é a especialidade médica que estuda o câncer e os tumores que podem ser desenvolvidos no organismo. Essa especialidade tem como principal função entender e encontrar formas de tratamento para o câncer, a fim de obter a melhora ou cura [71].

O câncer do colo do útero

O câncer do colo do útero, também é chamado de câncer cervical e ocorre quando células anormais se desenvolvem e se espalham no colo do útero, a parte inferior do útero. Geralmente é causado pela infecção persistente por alguns tipos do Papilomavírus Humano (HPV - Human Papilloma Virus), chamados de tipos oncogênicos. É transmitido durante a relação sexual.

A infecção genital por esse vírus é muito frequente e não causa doença na maioria das vezes. Entretanto, em alguns casos, ocorrem alterações celulares que podem evoluir para o câncer. Essas alterações são descobertas facilmente no exame preventivo (conhecido também como Papanicolau), e são curáveis na quase totalidade dos casos. Quando encontrado cedo, o câncer do colo do útero é altamente curável. Por isso, é importante a realização periódica desse exame [72].

Causas do câncer de colo de útero

O câncer do colo do útero costuma ser causado pelo vírus do papiloma humano (HPV), que é transmitido durante a relação sexual. Esse vírus também causa verrugas genitais. Os fatores de risco para o câncer do colo do útero incluem [72]:

- Início precoce da atividade sexual e múltiplos parceiros;

- Fumar cigarro;

- Uso prolongado de pílulas anticoncepcionais.

Diagnóstico do câncer de colo de útero

O diagnóstico de câncer de colo de útero é feito através de exames pélvico e história clínica, exame preventivo (Papanicolau), colposcopia e biópsia. Uma biópsia é realizada caso um tumor, ferida ou outra área anômala forem observados no colo do útero durante o exame pélvico ou caso displasia ou câncer sejam detectados pelo exame de Papanicolau.

Dois tipos distintos de exames são realizados: Biópsia por punção e Curetagem endocervical. Em biópsia por punção, um pedaço minúsculo do colo do útero, selecionado usando o colposcópio, é removido. Em curetagem endocervical, tecido que não pode ser visto é raspado do interior do colo do útero [72, 73].

5.2.2 O conjunto de dados

Um conjunto de dados, tanto aberto quanto público, foi utilizado para o trabalho.

Os dados foram obtidos do repositório UCI Machine Learning [1]. Esse conjunto de dados foi retirado do Hospital Universitário de Caracas em Caracas, Venezuela.

O conjunto de dados é formado por dados tabulares de câncer do colo do útero que contém indicadores e fatores de risco para prever se uma mulher terá câncer do colo do útero.

Dados tabulares são dados fornecidos em tabelas, com cada linha representando uma instância e cada coluna uma característica. As características incluem dados demográficos (como idade), estilo de vida e histórico médico. Esse conjunto de dados possui 858 (oitocentas e cinquenta e oito) instâncias e 36 (trinta e seis) atributos.

Neste caso, objetiva-se prever se um paciente tem câncer de colo de útero [73], com base em fatores ambientais e de diagnóstico.

5.2.3 Caso selecionado do conjunto de dados

Insta esclarecer que para melhor se compreender a explicação das técnicas de IA explicável locais, houve seleção de uma paciente do conjunto de dados que teve diagnóstico de ocorrência de câncer de colo de útero e com características atípicas na clínica médica.

O objetivo aqui era selecionar uma mulher com diagnóstico não muito simples para avaliação dos especialistas no domínio médico oncológico.

[1] https://archive.ics.uci.edu/ml/datasets.php

Idade
Num. parceiros sexuais
Idade 1º relação sexual
Num. gestações
Fuma (sim ou não)
Fuma(anos)
Fuma(maços/anos)
Contraceptivos hormonais (sim ou não)
Contraceptivos hormonais(anos)
DIU (Dispositivo Intrauterino) (sim ou não)
DIU (anos)
DSTs (Doenças sexualmente transmíssíveis):
condilomatose, condilomatose cervical, condilomatose vaginal,
condilomatose vulvo-perineal, sífilis, doença inflamatória pélvica,
herpes genital, molusco contagioso, AIDS, HIV, hepatite B e HPV
DSTs (quantidade)
DSTs: num de diagnósticos
DSTs: Tempo desde o primeiro diagnóstico
DSTs: Tempo desde o último diagnóstico
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Colposcopia
Teste de Schiller
Exame de Papanicolaou
Câncer (sim ou não)

Figura 17 – As características/atributos usados no modelo.

Assim, segue na tabela abaixo as características da paciente selecionada:

5.2.4 Escolha do modelo de aprendizagem de máquina

Nesta etapa foi criado um modelo de aprendizagem de máquina para classificação. O objetivo foi construir e interpretar o modelo preditivo para os resultados da biópsia do câncer do colo do útero com base nos resultados da citologia e outros fatores de risco em potencial, incluindo dados demográficos e histórico do paciente.

Como a biópsia serve como ponto fundamental para o diagnóstico de câncer cervical, então nos exemplos analisados, o resultado da biópsia (câncer) foi usado como variável preditora.

Na construção do modelo de aprendizagem de máquina foi utilizada a linguagem python, muito utilizada e popular em ciência de dados.

Os pacotes utilizados para a leitura dos dados, manipulação, tratamento inicial dos dados e treinamento foram o numpy [2] e pandas[3]. Já as modelagens de predição foram feitas com o pacote ScikitLearn [4].

[2] https://numpy.org
[3] https://pandas.pydata.org
[4] https://scikit-learn.org/stable/

Tabela 2 – Informações da paciente selecionada

Idade: 21 anos
Sua primeira relação sexual ocorreu aos 15 anos
Ela teve 4 parceiros sexuais em sua vida
Teve 1 gravidez
Não é fumante
Não faz uso de contraceptivos hormonais
Não faz uso do contraceptivo DIU (dispositivo intra-uterino)
Não possui nenhum tipo de DSTs (Doença Sexualmente transmissíveis), como: Condiloma, Condiloma cervical, Condiloma vaginal, Condiloma vulvo-perineal, sífilis, doença inflatória pélvica, herpes genital, molusco contagioso, AIDS, HIV, hepatite B e HPV
Exame de colposcopia
Teste de Schiller: Positivo (toda vez que houver alguma área amarelada do colo uterino, que não fica corada com o lugol, sugerindo a presença de células atípicas).
Ela realizou o exame citopatológico do colo do útero (ou exame de Papanicolau)
Obs: A coleta periódica do exame citopatológico do colo do útero (ou exame de Papanicolau) possibilita o diagnóstico precoce, tanto das formas pré-cancerosas, como do câncer propriamente dito. No exame ginecológico rotineiro, além da coleta do material citopatológico, é realizado o Teste de Schiller (coloca-se no colo do útero uma solução iodada) para detectar áreas não coradas, suspeitas. A colposcopia (exame em que se visualiza o colo do útero com lente de aumento de 10 vezes ou mais) auxilia na avaliação de lesões suspeitas ao exame rotineiro, e permite a realização de biópsia dirigida (coleta de pequena porção de colo do útero), fundamental para o diagnóstico de câncer [73].

As etapas da construção do modelo foram: coleta de dados, preparação dos dados e escolha do modelo.

1. Coleta de dados

O conjunto de dados foi disponibilizado pelo UCI - Machine Learning Repository. O conjunto de dados contém informações demográficas, hábitos e histórico médico de 858 (oitocentos e cinquenta e oito) pacientes e 36 (trinta e seis) características.

É importante mencionar que não foi realizada nenhuma restrição na seleção dessas instâncias, ou seja, todas as instâncias (858 pacientes) e características (36) foram usadas.

Há uma particularidade de que todos os pacientes são do sexo feminino.

Na sequência, fez-se uma análise exploratória dos dados para uma melhor visualização das variáveis.

2. Preparação dos dados

Para o pré-processamento dos dados, efetuou-se o método de imputação de dados de média ponderadas para o tratamento dos valores faltantes e dos outliers e, para a normalização, realizou-se o procedimento de escala por máximos e mínimos. Após a verificação de inconsistência e limpeza do banco de dados, dividiu-se 75% dos dados para treinamento e 25% para teste.

3. Construindo e treinando o modelo

Nesta etapa, aplicou-se 07 (sete) diferentes algoritmos: 1) Naive bayes, 2) random forest (floresta aleatótia), 3) regressão logística, 4) máquina de vetores de suporte (SVM), 5) árvore de decisão, 6) rede neural multicamadas (MLP) e 7) KNN. O critério de decisão para a escolha do algoritmo foi aquele de maior exatidão (accuracy) com os dados de teste e algoritmo de caixa-preta.

Neste caso, usou-se alguns algoritmos simples como, por exemplo, árvores de decisão, mas apenas com o intuito de testar. O algoritmo selecionado foi o método ensemble Random Forest (Floresta aleatória).

Neste contexto, a figura 18 a seguir exibe os dados de exatidão do modelo selecionado e a tabela 4 exibe os dados de exatidão para os outros modelos.

Após a seleção do algoritmo (Floresta aleatória), foi construído um programa para aplicar as técnicas de explicação. Os critérios para a seleção das técnicas de IA explicável são descritos abaixo.

5.3 Escolha das técnicas de inteligência artificial explicável

O presente estudo optou por selecionar técnicas de IA explicável independentes de modelo (agnósticas), haja vista a flexibilidade para usar qualquer modelo, bem como qualquer explicação e, ainda, devido ao menor custo de mudança de modelo.

Outra premissa adotada na escolha das técnicas era a utilização de técnicas populares e consideradas estado da arte atualmente.

Os métodos de IA explicável SHAP, LIME e Permutation Importance foram escolhidos como escopo deste trabalho, já que são técnicas independentes de modelo (model-agnostic) e são técnicas populares.

```
# Utilizando um classificador RandomForest
from sklearn import metrics

from sklearn.ensemble import RandomForestClassifier
modelo_v2 = RandomForestClassifier(random_state = 42)
modelo_v2.fit(X_treino, Y_treino.ravel())

print("================================================================================")
# Verificando os dados de treino
rf_predict_train = modelo_v2.predict(X_treino)
print("Exatidão (Accuracy) com os dados de treino do modelo RandomForest: {0:.4f}".format(metrics.accura
# Verificando nos dados de teste
rf_predict_test = modelo_v2.predict(X_teste)
print("Exatidão (Accuracy) com os dados de teste do modelo RandomForest: {0:.4f}".format(metrics.accuracy
print()
print("Confusion Matrix - Modelo RandomForest")
print("{0}".format(metrics.confusion_matrix(Y_teste, rf_predict_test, labels = [1, 0])))
print()
print("Classification Report RandomForest")
print(metrics.classification_report(Y_teste, rf_predict_test, labels = [1, 0]))
print("================================================================================")
```

```
================================================================================
Exatidão (Accuracy) com os dados de treino do modelo RandomForest: 0.9917
Exatidão (Accuracy) com os dados de teste do modelo RandomForest: 0.9496

Confusion Matrix - Modelo RandomForest
[[  4  11]
 [  2 241]]

Classification Report RandomForest
              precision    recall  f1-score   support

           1       0.67      0.27      0.38        15
           0       0.96      0.99      0.97       243

   micro avg       0.95      0.95      0.95       258
   macro avg       0.81      0.63      0.68       258
weighted avg       0.94      0.95      0.94       258
```

Figura 18 – Algoritmo Randon Forest. Exatidão (Accuracy) com os dados de teste de: 09496

5.4 Método de coleta de dados

O método de coleta de dados selecionado foi o Método de Explicitação do Discurso Subjacente (MEDS). Registra-se que MEDS é um método qualitativo que têm como objetivo principal ouvir detalhadamente aquilo que os entrevistados têm a dizer, em contextos naturais e da forma mais livre possível [74].

O MEDS é um método exploratório que deriva de perguntas abertas e não de perguntas prontas. Ele é adequado à pesquisa daquilo que é desconhecido, ou seja, onde se tem pouco conhecimento prévio. As entrevistas (coleta de dados) são realizadas em encontros presenciais que têm como modelo as conversas cotidianas em contextos informais [70,75].

As fases do MEDS são:

• Fase 1: Seleção da amostra

• Fase 2: A Construção do roteiro para as entrevistas

Tabela 3 – Tabela de Modelos x Exatidão (Accuracy)

Modelo	***Exatidão (Accuracy)***
Naive Bayes	Exatidão (Accuracy): 0.7713 precision: 0.94 recall: 0.77 f1-score: 0.83
Regressão Logística	Exatidão (Accuracy): 0.9574 precision: 0.95 recall: 0.96 f1-score: 0.96
Árvore de Decisão	Exatidão (Accuracy): 0.9496 precision: 0.95 recall: 0.95 f1-score: 0.95
Máquina de vetores de suporte (SVM)	Exatidão (Accuracy): 0.9496 precision: 0.94 recall: 0.94 f1-score: 0.94
Rede Neural Multicamadas (MLP)	Exatidão (Accuracy): 0.9412 precision: 0.89 recall: 0.94 f1-score: 0.91
KNN	Exatidão (Accuracy): 0.9341 precision: 0.89 recall: 0.93 f1-score: 0.91

- Fase 3: As entrevistas
- Fase 4: A Transcrição dos depoimentos
- Fase 5: A Análise dos depoimentos coletados

5.4.1 Fase 1: Seleção da amostra

o campo da pesquisa qualitativa também usa o conceito de amostra. O MEDS busca uma homogeneidade e privilegia o recrutamento que é denominado “perfil de alta definição” [70].

Recrutamento dos participantes

O recrutamento dos participantes para a pesquisa visa profissionais de saúde com experiência em qualquer fase da oncologia, seja com pesquisa, diagnóstico, consulta médica,

tratamento ou pós-tratamento.

Decisões a respeito do tamanho da amostra

O MEDS não estipula um número exato de participantes, sendo o principal critério para determinar se as entrevistas são suficientes a saturação.

A saturação ocorre quando o entrevistador começa a ouvir relatos muito semelhantes àqueles que já ouviram [70]. Neste estudo, houve um total de 12 (doze) entrevistas.

5.4.2 Fase 2: A Construção do roteiro para as entrevistas

O MEDS tem um guia de construção de roteiro detalhista e estipula algumas diretrizes gerais para a construção de roteiros.

As principais diretrizes são [74]:

(a) o roteiro deve ser estruturado em sua concepção e flexível em sua aplicação;

(b) o roteiro deve se inspirar em conversas naturais;

(c) o roteiro deve constar apenas de itens essenciais para as próprias entrevistas, a fim de evitar que as perguntas sejam simplesmente lidas e soem artificiais para os entrevistados;

(d) para que o entrevistador possa conhecer o ponto de vista do entrevistado, é necessário gerar perguntas abertas que comportem qualquer tipo de resposta e perguntas de esclarecimento e/ ou aprofundamento também devem ser previstas;

(e) para preservar a naturalidade de uma conversa informal, alguns itens deverão gerar perguntas fechadas, seguidas de perguntas de esclarecimento e/ ou aprofundamento;

(f) itens que geram perguntas que solicitam opiniões, reflexões, posturas, sentimentos, avaliações etc. do entrevistado – a respeito de determinados tópicos devem poder ser confrontados com itens que geram informações objetivas a respeito dos mesmos tópicos.

Houve a criação de 02 (dois) grupos de perguntas para o questionário.

As perguntas do primeiro grupo são relacionadas ao entendimento e interpretação dos resultados (saídas) das técnicas de IA explicável e elas são feitas para todos as técnicas de IA explicável avaliados (SHAP, LIME e Permutation Importance).

Já as perguntas do segundo grupo são relacionadas à percepção de confiança e ao uso

da inteligência artificial no domínio médico.

Segue abaixo a lista das perguntas utilizadas no questionário:

I) Primeiro grupo de perguntas

• Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pela técnica XYZ?

• Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

• O quanto você concorda com a seguinte afirmação: "A explicação gerada pela técnica XYZ é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer de colo de útero da paciente?". Por quê?

• Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo XYZ? Que informação seria interessante ser apresentada nesse método?

II) Segundo grupo de perguntas

• Qual técnica de explicação você mais gostou? Por quê?

• Você confia na inteligência artificial/aprendizagem de máquina? Por quê?

• Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

• Que sugestão ou crítica você teria para que essas técnicas de IA explicável sejam implantados no domínio médico (oncologia)?

5.4.3 Fase 3: As entrevistas

No MEDS, cada participante é entrevistado individualmente por um único entrevistador e é realizada uma única entrevista com cada participante.

As entrevistas acontecem em horários negociados entre as partes e geralmente tem duração de, no máximo, 01 (uma) hora.

As entrevistas são gravadas na íntegra (em áudio) com o consentimento dos entrevistados.

As entrevistas ocorrem dentro de um modelo de conversa informal e em lugares em

os participantes se sentem à vontade, familiarizados [70, 74]. Os entrevistados optaram por realizar as entrevistas em seus respectivos locais de trabalho.

Antes da entrevista, os participantes assinaram um termo de livre consentimento do qual constam informações sobre os objetivos da pesquisa, bem como sobre os eventuais riscos que ela pode representar para aqueles que dela participam e sobre o uso que pode ser feito do material coletado.

A realização das entrevistas

O MEDS estipula que o entrevistador deverá ter em mãos um roteiro estruturado que deverá ser aplicado de forma bem flexível, ou seja, a ordem dos itens pode ser alterada, dependendo dos pronunciamentos dos entrevistados. E em caso de um ou mais itens serem abordados de forma espontânea pelo entrevistado, então não há necessidade de o mesmo ser transformado em perguntas. O MEDS incentiva a introdução espontânea somente de perguntas de aprofundamento ou esclarecimento [70, 74].

As entrevistas foram realizadas com 12 (doze) profissionais com idade entre 35 e 60 anos. Os profissionais entrevistados trabalham em 6 (seis) lugares distintos: a) INCA - Instituto Nacional de Câncer, b) Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz / Fiocruz, c) Ambulatório São Lucas do Departamento de Medicina da PUC-Rio, d) Consultório clínico particular, e) Instituto Nacional da Saúde da Mulher, da Criança e do Adolescente-Fernandes Figueira / Fiocruz e f) Hospital Oncologia D'Or Barra.

No que tange à formação acadêmica básica, temos a seguinte distribuição: 7 (sete) profissionais com graduação em medicina; 01 (um) profissional com graduação em ciências biológicas; 2 (dois) profissionais com graduação em ciências biológicas - modalidade médica; 1 (um) profissional com graduação em ciências biológicas - modalidade genética e 1 (um) profissional com graduação em farmácia.

Dos 12 (doze) entrevistados, temos 4 (quatro) homens e 8 (oito) mulheres e para manter o anonimato dos voluntários, seus nomes foram codificados em PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, PO9, PO10, PO11 e PO12 (profissionais de oncologia).

A figura 19, apresenta o perfil dos entrevistados.

Os entrevistados selecionados foram recrutados a partir de contato pessoal ou telefônico dos pesquisadores, sendo certo que eles integravam parte do ciclo social e profissional dos pesquisadores. Vale dizer que o recrutamento dos participantes foi uma das maiores dificuldades do presente estudo.

Código	Sexo	Local trabalho	Graduação
PO01	M	INCA - Instituto Nacional de Câncer	Medicina
PO02	F	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz	Ciências biológicas
PO03	F	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz	Ciências biológicas - modalidade médica
PO04	F	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz	Ciências biológicas - modalidade médica
PO05	F	Ambulatório São Lucas do Departamento de Medicina da PUC-Rio	Medicina
PO06	F	Consultório clínico particular	Medicina
PO07	F	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz	Farmácia
PO08	M	Instituto Nacional da Saúde da Mulher, da Criança e do Adolescente-Fernandes Figueira / Fiocruz	Medicina
PO09	F	Instituto Nacional da Saúde da Mulher, da Criança e do Adolescente-Fernandes Figueira / Fiocruz	Medicina
PO10	M	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz	Ciências biológicas - modalidade genética
PO11	F	Hospital Oncologia D'Or Barra	Medicina
PO12	M	Hospital Oncologia D'Or Barra	Medicina

Figura 19 – Perfil dos entrevistados. Fonte: Coleta de dados

Material utilizado durante as entrevistas

Como material de apoio para as entrevistas foram desenvolvidos: 1) uma apresentação pré-entrevista sobre o objetivo da pesquisa, além de conceitos relacionados à inteligência artificial, aprendizagem de máquina e IA explicável e 2) um termo de consentimento com orientações sobre a entrevista e algumas considerações éticas.

As entrevistas foram realizadas no local de trabalho dos participantes, em um laboratório móvel, constituído por notebook com todo o script aberto com os resultados das técnicas de IA explicável no ambiente jupyter notebook [5].

O questionário sempre foi feito com base na mostra dos resultados em tempo real. A captura de voz nas entrevistas foi feita com o aplicativo de gravação de voz do celular do pesquisador.

Antes do início das entrevistas, os usuários foram orientados sobre os procedimentos e a eles foi demonstrada uma apresentação do objetivo da pesquisa e conceitos relacionados. Na sequência, o termo de consentimento foi entregue para leitura e assinatura do usuário.

Além disso, os participantes foram informados de que o papel do pesquisador era apenas de orientar, realizar as perguntas e acompanhar as respostas, de modo que ele não

[5] https://jupyter.org

poderia ajudá-los durante a entrevista e que o objetivo era avaliar as técnicas de IA explicável e não o participante, para que o mesmo não se sentisse intimidado ou envergonhado com eventuais dificuldades nas respostas.

A figura 20 apresenta informações detalhadas das entrevistas realizadas.

Código	Data Entrevista	Tempo Entrevista	Local Entrevista
PO01	27/12/2019	20:59	INCA - Instituto Nacional de Câncer
PO02	27/01/2020	09:50	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz
PO03	27/01/2020	23:09	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz
PO04	27/01/2020	23:26	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz
PO05	30/01/2020	20:22	Ambulatório São Lucas do Departamento de Medicina da PUC-Rio
PO06	03/02/2020	18:54	Consultório clínico particular
PO07	03/02/2020	20:59	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz
PO08	10/02/2020	18:03	Instituto Nacional da Saúde da Mulher, da Criança e do Adolescente-Fernandes Figueira / Fiocruz
PO09	11/02/2020	25:38	Instituto Nacional da Saúde da Mulher, da Criança e do Adolescente-Fernandes Figueira / Fiocruz
PO10	12/02/2020	16:53	Laboratório de Genômica Funcional e Bioinformática IOC / Fiocruz
PO11	13/02/2020	21:33	Hospital Oncologia D'Or Barra
PO12	13/02/2020	25:07	Hospital Oncologia D'Or Barra

Figura 20 – Entrevistas. Fonte: Coleta de dados

5.4.4 Fase 4: A transcrição dos depoimentos

O MEDS aconselha que o nível de detalhamento das transcrições seja pensado e descrito caso a caso e também enfatiza que as falas dos entrevistados não devem ser alteradas ou editadas. As entrevistas foram gravadas e transcritas na íntegra pelo próprio pesquisador para serem utilizadas na análise dos dados, conforme apresentadas no Apêndice B.

5.4.5 A análise dos depoimentos coletados

Para a análise dos depoimentos coletados foi utilizada a Teoria Fundamentada em Dados (Grounded Theory).

A Teoria Fundamentada em Dados será detalhada na próxima seção.

5.5 Método de Análise Qualitativa

A Teoria Fundamentada nos Dados gera explicações, com a mínima intervenção do pesquisador, sobre a ação dos indivíduos em um contexto delimitado e a partir da realidade deles [69].

A Teoria Fundamentada nos Dados é baseada na ideia de codificação (coding), que é o processo de analisar os dados. Os dados revelam o comportamento do indivíduo em face de situações específicas [69].

Durante a codificação são identificados conceitos (ou códigos) e categorias.

Um conceito (ou código) dá nome a um fenômeno de interesse para o pesquisador; abstrai um evento, objeto, ação, ou interação que tem um significado para o pesquisador [69].

Categorias são agrupamentos de conceitos unidos em um grau de abstração mais alto.

Codificar não significa meramente associar trechos do texto a códigos ou categorias, mas sim fazer questionamentos e dar respostas provisórias sobre categorias e suas relações que são verificadas e aperfeiçoadas ao longo das 03 (três) fases do processo de codificação: codificação aberta, codificação axial e codificação seletiva [69].

5.5.1 Codificação aberta (open coding)

A primeira etapa do método Teoria Fundamentada nos Dados (codificação aberta) envolve a quebra, a análise, a comparação, a conceituação e a categorização dos dados.

Em outras palavras, envolve a criação de códigos que estão relacionados a trechos do texto relevantes para a pesquisa, conforme exemplificado na Figura 21.

Nas fases iniciais da codificação aberta, o pesquisador explora os dados examinando minuciosamente aquilo que lhe parece relevante devido à leitura intensiva dos textos.

Na fase de codificação aberta os incidentes ou eventos são agrupados em códigos através da comparação teórica [69].

O processo comparativo durante a codificação é o processo central de análise da Teoria Fundamentada nos Dados (TFD).

As comparações teóricas são feitas nas fases iniciais do processo de pesquisa ou quando algo novo surge nos dados e contribuem para a identificação de categorias con-

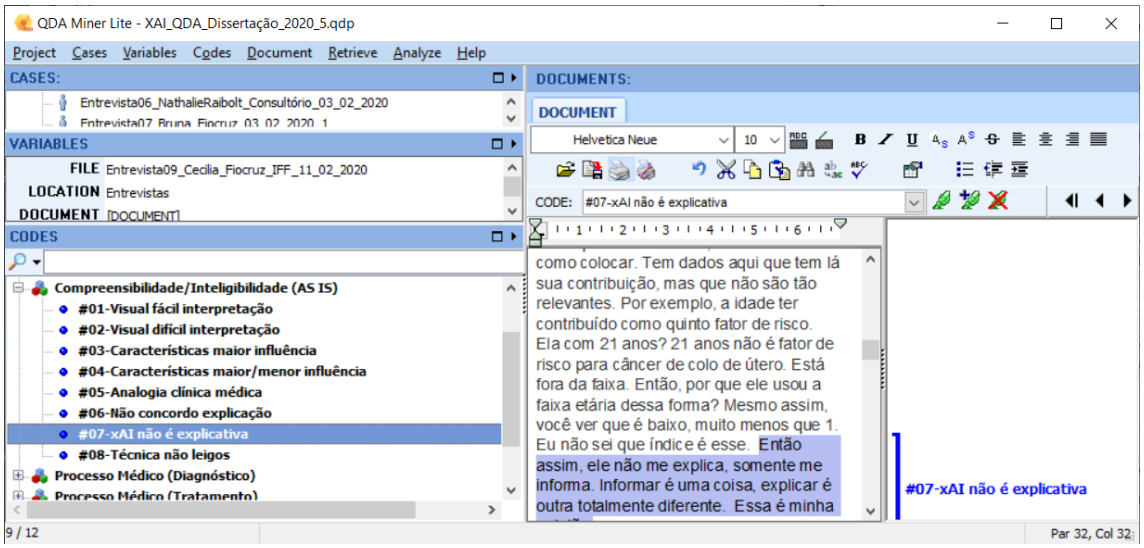

Figura 21 – Etapa de codificação aberta (Open Coding) com a ferramenta Qda Miner

ceituais.

Várias interações de comparações foram realizadas para a seleção de códigos que indicavam relatos representativos em citações no texto.

Através da codificação aberta foi encontrado um conjunto de códigos que podem representar possíveis entendimentos da percepção dos especialistas no domínio quanto às técnicas de inteligência artificial explicável e sobre a inteligência artificial.

Após várias revisões dos códigos gerados com outro pesquisador envolvido na pesquisa, foi possível identificar padrões para a criação e classificação de categorias. Foram encontrados 17 (dezessete) códigos relacionados às técnicas de IA explicável e 15 (quinze) códigos relacionados à inteligência artificial no domínio médico.

A ferramenta QDA Miner foi utilizada em todo o processo de análise qualitativa dos dados. A ferramenta QDA Miner [6] é um software qualitativo de análise de dados para organizar, codificar, anotar, recuperar e analisar coleções de documentos e imagens.

Essa ferramenta pode ser usada para analisar transcrições de entrevistas ou grupos focais, documentos legais, artigos de periódicos, discursos e até livros inteiros, bem como desenhos, fotografias, pinturas e outros tipos de documentos visuais.

[6] https://provalisresearch.com/products/qualitative-data-analysis-software/freeware

5.5.2 Codificação axial (axial coding)

O objetivo da codificação axial é iniciar o processo de reagrupamento dos dados que foram divididos durante a codificação aberta. Na codificação axial, as categorias são relacionadas às suas subcategorias para gerar explicações mais precisas e completas sobre os fenômenos. A meta é desenvolver sistematicamente as categorias e relacioná-las. Esse passo da análise é importante para construir a teoria.

Os 32 (trinta e dois) códigos foram analisados e agrupados de acordo com suas propriedades, formando assim conceitos que representam categorias. Na etapa de codificação axial foram criadas 4 (quatro) categorias:

a) explicabilidade das técnicas de IA explicável;

b) melhoria da explicabilidade;

c) confiança na inteligência artificial e

d) recomendação para a implantação de técnicas de IA explicável no domínio médico.

Neste tocante, para a melhor compreensão do leitor, cabe apresentar a definição das categorias de análise acima citadas, senão vejamos:

a) **Explicabilidade das Técnicas de IA Explicável**:

Essa categoria foi assim nomeada, pois contém informações ou códigos relacionados ao entendimento atual das técnicas de explicação. Essa é a principal categoria relacionada ao entendimento das técnicas de inteligência artificial explicável.

b) **Melhoria da Explicabilidade das Técnicas de IA Explicável**:

Essa categoria surgiu a partir da necessidade de agrupar os códigos de sugestão de melhoria das técnicas de IA explicável. Esses códigos se referem aos problemas ou fragilidades, bem como as oportunidades de melhoria dos métodos de explicação.

c) **Confiança na inteligência artificial**:

Essa categoria foi criada com o objetivo de entender se os usuários confiam na inteligência artificial/aprendizagem de máquina.

d) **Recomendação para a implantação de técnicas de IA explicável no domínio médico**:

Essa categoria foi criada com o objetivo de entender as recomendações para a implan-

tação de técnicas de IA explicável no domínio médico.

Inicialmente, cumpre apresentar os códigos e suas respectivas definições utilizados nas pesquisas aplicando a categoria de Explicabilidade das Técnicas de IA Explicável, a categoria Melhoria da Explicabilidade das Técnicas de IA Explicável, a categoria Confiança na inteligência artificial e a categoria Recomendação para a implantação de técnicas de IA explicável no domínio médico.

Tabela 4 – Tabela com códigos e descrições da categoria Explicabilidade das Técnicas de IA Explicável:

Códigos	***Descrição do código***
#01-Visual de fácil interpretação	Esse código está relacionado ao entendimento do usuário especialista no domínio de que o resultado da técnica xAI é de fácil interpretação.
#02-Visual é de difícil interpretação	Esse código está relacionado ao entendimento do usuário especialista no domínio de que o resultado da técnica xAI é de difícil interpretação.
#03-Características com maior influência	Esse código está relacionado ao entendimento do usuário especialista no domínio da explicação da técnica de XAI. Nesse caso, as características são enumeradas em ordem de maior influência no resultado da explicação.
#04-Características com maior/menor influência	Esse código está relacionado ao entendimento do usuário especialista no domínio da explicação da técnica de XAI. Nesse caso, as características são apresentadas mostrando as de maior influência e menor influência na explicação.
#05-Analogia clínica médica	Esse código está relacionado ao entendimento do usuário especialista no domínio que a apresentação da técnica de XAI é análoga à clínica médica.
#06-Não concordo com o resultado	Esse código está relacionado ao entendimento do usuário especialista no domínio de que os resultados apresentados pela técnica de XAI não condizem com os resultados da clínica médica.
#07-xAI não é explicativa	Esse código está relacionado ao entendimento do usuário especialista no domínio de que os resultados apresentados pela técnica de XAI não são explicativos.
#08-Técnica não é para leigos	Esse código está relacionado ao entendimento do usuário especialista no domínio de que os resultados apresentados pela técnica de XAI não são apropriados para um usuário leigo.

Fonte: Elaborado pelo autor

Tabela 5 – Tabela com códigos e descrições da categoria Melhoria da Explicabilidade das Técnicas de IA Explicável

Códigos	***Descrição do código***
#01. Detalhar características	Esse código está relacionado ao fato de que algumas características importantes para o domínio do câncer de colo de útero (Colposcopia, Teste de Schiller e Exame de Papanicolau) e portanto muito usado como fator de decisão na clínica médica, serem características binárias (ou seja, "é S ou é N").
#02-Explicar como interpretar	Esse código está relacionado a sugestão de incluir um guia ou explicação de como a técnica deve ser interpretada, com o significado de cada símbolo, cores do gráfico, números e suas respectivas medidas.
#03-Novos tipos visualização gráfica	Esse código está relacionado à sugestão de disponibilizar outros tipos de gráficos para que o usuário possa escolher de acordo com sua experiência.
#04-Mostrar todas as características da predição	Esse código está relacionado à sugestão de mostrar todas as características (features) utilizadas na predição do modelo.
#05. Incluir peso de contribuição das características	Esse código, significa incluir de maneira clara o peso, ou seja, a contribuição de cada característica na explicação.
#06-Alterar codificação visual	Alterar a codificação visual está relacionado à possibilidade de alterar tamanho, cores, saturação das cores, formato e textura das saídas das técnicas de IA explicável.
#07-Agrupar características por peso	Agrupar as caraterísticas pelo peso, ou seja, criar uma hierarquia de grupos de características (maior influência, influência intermediária e menor influência) para classificar os resultados.
#08-Novos tipos de visualização de dados	Esse código significa incluir outras formas de visualização de dados, como: mapas mentais, organogramas e modelos matemáticos.
#09-Incluir rastreabilidade da informação	Refere em mostrar os relacionamentos existentes entre o conjunto de dados, o processamento da técnica de IA explicável e sua respectiva saída no resultado (gráfico).

Fonte: Elaborado pelo autor

Tabela 6 – Tabela com códigos e descrições da categoria Confiança em Inteligência Artificial

Códigos	***Descrição do código***
#01-Testar/validar	Esse código se refere à estudos, ter as técnicas de IA validadas e testadas por um número grande de pessoas, áreas e setores. Envolve também comparar os resultados com as práticas atuais, a replicabilidade ou reprodutibilidade, ou seja, envolve a medida em que um processo de tomada de decisão por IA pode ser repetido com o mesmo resultado.
#02-Considerações éticas	Esse código está relacionado ao uso da inteligência artificial de maneira inadequada, ou seja, sem levar em consideração questões éticas.
#03-Curadoria de dados	Curadoria de dados envolve uma série de atividades voltadas para a gestão de dados. Inclui planejamento, criação, seleção de formatos e documentação do conjunto de dados.
#04-Maior assertividade	Esse código está relacionado ao entendimento de que a inteligência artificial aumenta a assertividade.
#05-Não substitui o profissional de saúde	Esse código está relacionado ao entendimento de que a inteligência artificial não substitui as atividades do profissional de saúde.
#06-Uso IA como apoio	Esse código está relacionado ao entendimento de que a inteligência artificial serve como apoio ao processo decisório e não como decisão final.
#07-Viés algorítmico	Esse código está relacionado ao receio do viés algorítmico influenciar negativamente os resultados da inteligência.
#08-Conhecer limitadores da tecnologia	Esse código se refere a conhecer os pontos fracos e fatores de sucesso das técnicas de inteligência artificial. Ter um entendimento de que fatores contribui para um erro ou acerto.
#09-Treinar profissionais	Esse código se refere a qualificação profissional. A evolução das soluções de IA é diária e um desafio para os profissionais de IA é desenvolver soluções e ao mesmo manter-se atualizado com as novidades que surgem quase que diariamente neste campo.
#10-Big data é facilitador	"Esse código se refere ao entendimento de que a grande quantidade de dados disponíveis atualmente é um facilitador para o aumento de confiança na IA.
#11-Necessidade do fator humano	Esse código se refere ao entendimento da necessidade do fator humano para à confiança na IA.

Tabela 7 – Tabela com códigos e descrições da categoria Recomendações para a Implantação de Técnicas de IA Explicável no Domínio Médico

Códigos	***Descrição do código***
#01 - Popularizar as técnicas de IA explicável	Esse código significa anunciar, difundir e divulgar as potencialidades e resultados da inteligência artificial e das técnicas de IA explicável no contexto médico.
#02 - Participação do especialista do domínio	Esse código está relacionado com a sugestão de incluir um especialista no domínio em todo o ciclo de vida.
#03 - Selecionar, usar e testar diversos conjunto de dados	Esse código está relacionado ao conceito de que o conjunto de dados ser um importante insumo no processo de aprendizagem de máquina.
#04-Evidência da confiança e aplicabilidade do modelo	Esse código está relacionado as ações para mitigar os problemas de desconfiança em informações e decisões geradas pelo aprendizagem de máquina.

5.5.3 Codificação seletiva (selective coding)

A codificação seletiva refina todo o processo identificando a categoria central (core category) da teoria, que deve ser capaz de integrar todas as outras categorias e expressar a essência do processo social. Esta categoria central pode ser criada ou pode ser uma categoria existente [69].

A categoria central na pesquisa é explicabilidade das técnicas de IA explicável, uma vez que se busca entender se as técnicas de IA explicável oferecem explicabilidade aos especialistas do domínio.

5.5.4 Avaliação dos resultados

Na etapa de avaliação dos resultados, os códigos e os relacionamentos foram analisados em conjunto com outro pesquisador que é especialista em metodologia qualitativa, de forma a verificar as análises realizadas.

É verdade que várias interpretações podem coexistir, portanto o papel deste pesquisador foi se certificar de que a codificação, categorias e análise foram desenvolvidas de acordo com os procedimentos do método.

Foram realizadas aproximadamente 10 (dez) interações presenciais e online para alcançar os resultados esperados.

6. Resultados da Avaliação Humana das Técnicas de Inteligência Artificial Explicável

Este capítulo tem por finalidade avaliar os resultados da explicabilidade das técnicas de inteligência artificial explicável, bem como da confiança na inteligência artificial sob a perspectiva do usuário especialista no domínio médico. Deste modo, as categorias objeto de análise são:

a) explicabilidade das técnicas de IA explicável;

b) melhoria da explicabilidade;

c) confiança na inteligência artificial e

d) recomendação para a implantação de técnicas de IA explicável no domínio médico.

6.1 Resultados da explicabilidade das técnicas de IA explicável

Nesta seção são apresentados os resultados da Explicabilidade das Técnicas de IA explicável e melhoria da explicabilidade para as técnicas SHAP, LIME e Permutation Importance.

6.1.1 Resultados da explicabilidade da técnica SHAP

Este tópico destina-se a apresentar os resultados de explicabilidade da técnica SHAP. A figura 22 exibe as explicações visuais dos fatores de risco para o câncer de colo de útero utilizando-se a técnica SHAP, como se vê na imagem abaixo.

Assim, temos na parte vermelha do gráfico as características que mais influenciam na previsão de câncer e, na parte azul, as características que menos influenciam na previsão de câncer.

Como se vê, no caso em análise, a mulher em questão tem um alto risco previsto de 0,91, já que os principais fatores para o risco de câncer foram: o teste de schiller = sim (1); exame de papanicolau = sim (1); num.parceiro sexuais = 4; idade = 21; idade 1º relação sexual = 15.

É importante destacar que as características do teste de schiller e do exame de papanicolau são binárias (ou seja, "é S ou é N").

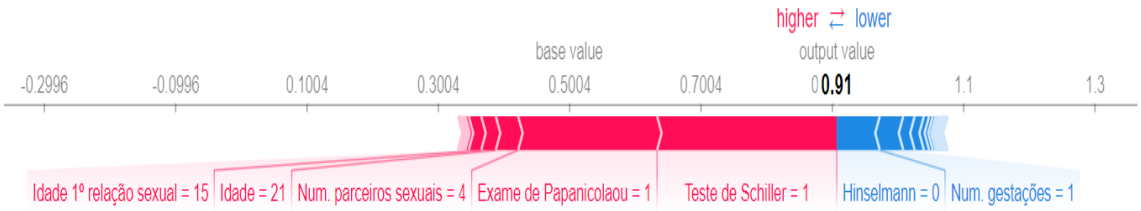

Figura 22 – Técnica SHAP. Fatores de risco para câncer de colo do útero

Desta forma, é importante demonstrar na tabela 8 os resultados da categoria Explicabilidade das Técnicas de IA Explicável para a técnica SHAP:

Tabela 8 – Resultados da categoria explicabilidade das técnicas de IA explicável - SHAP

Códigos e frequência	***Exemplo de Opinião***
Visual de fácil interpretação (9 entrevistas)	"... mas que de uma forma geral ficou bem claro o que é levado em conta para chegar nesse resultado de 91 %."(PO12).
Características com maior influência (12 entrevistas)	"É, o que me parece é que os dois testes e exames de rastreamento né, tanto o Schiller quanto o Papanicolau são bastante importantes para predizer a predisposição de ter ou não esse tipo de câncer. O que eu acho muito válido e relevante, ele ter dado esses 2 como os principais, depois ele vem com o número de parceiros sexuais e com a idade da primeira relação sexual."(PO04).
Características com maior/menor influência (5 entrevistas)	"A visualização rápida do que realmente é relevante em termos de fatores predisponentes para câncer e doenças pré-câncer. Num único olhar é possível ver graficamente o que é mais relevante ou não."(PO05).
Não concordo com o resultado (1 entrevistas)	"Esse caso específico, eu entendi o resultado, mas não faz muito sentido esse resultado. Eu não sei como ele chegou a esse número, não é uma tendência que a gente ver na prática clínica."(PO06).

xAI não é explicativa (2 entrevistas)	"Ele mostrou as contribuições de uma maneira mais gráfica que os outros. Mas ele não explica, ele só me diz os parâmetros que ele utilizou usou e em que percentual ele utilizou esses parâmetros."(PO09).
Técnica não é para leigos (2 entrevistas)	"E ai eu acho que tem a ver com o background da pessoa, ou seja, tem a ver com expertise de quem está avaliando. Então, para um especialista, um oncologista ou ginecologista, eu acho mais fácil de entender do que um público leigo. Porque um paciente olhando isso aqui, o número de parceiros, o exame, para ele não fica muito claro. Mas quem entende um pouco da doença, consegue identificar bem, então para o especialista está excelente."(PO02).

Por outro lado, a tabela 9 a seguir exibe os resultados da categoria Melhoria da Explicabilidade das Técnicas de IA Explicável para a técnica SHAP:

Tabela 9 – Resultados da categoria Melhoria da explicabilidade das técnicas de IA explicável - SHAP

Códigos e frequência	***Exemplo de Opinião***
Detalhar características (5 entrevistas)	"Eu acho que facilitaria dizer qual o tipo de alteração tem nesse exame de Papanicolau. Porque só tem câncer de colo de útero em quem apresenta doenças pre câncer do colo. Por exemplo uma lesão de baixo grau ou uma lesão mais especifica ela não fala tão a favor para uma doença do colo de útero, eu não levaria tanto em consideração, como se fosse uma lesão de alto grau ou outras alterações. Como eu te falei, a presença de captura hibrida, Teste de Schiller, dependendo da associação com a Colposcopia eu levaria mais em consideração o exame de Papanicolau. O que que deu nesse exame? Qual foi o resultado dele? Não sei o que deu aí, qual foi alteração? Tem uma gama de coisas que podem acontecer e nem tudo significa doenças pré câncer. Isso deixaria um pouco mais confiável."(PO05).
Explicar como interpretar (3 entrevistas)	"Aí você coloca o que é maior predição, baixa predição, de uma forma clara, para uma pessoa leiga entender, ela fica perfeita."(PO03).

Mostrar todas as características (2 entrevistas)	"Que ai não aparece todas as variáveis usadas na predição. Ai aparecem pelo visto as mais importantes, tanto para predição de doença, quanto para a ausência delas. No modelo anterior(LIME), lista todas elas. A vantagem do outro é que mostra todas as variáveis, ai fica claro dever o que está influenciando e o que não está."(PO08).
Incluir peso das características (2 entrevistas)	"...você tem uma informação que eu estou imaginando o tamanho da barra, diz que é de 0,4 a 30 e poucos por cento tá prevendo a partir do teste de Schiller. Talvez botar aqui o percentual que cada um contribuiu para essa predisposição no final. Acho que ficaria legal desse 81%, tantos por cento são advindos do lado de teste de Schiller, tantos por tanto do outro e 10% são desse. Aí você tem como dizer o quanto que o número de parceiros sexuais por exemplo, ou de cada um deles, tá contribuindo para o resultado final. Acho que seria interessante, ter de alguma forma, talvez aqui embaixo, um valor em percentual."(PO04).
Agrupar características (1 entrevista)	"O tipo de resultado, pelo menos agrupado. Agrupado com alto risco ou baixo risco para câncer."(PO06).
Incluir rastreabilidade da informação (3 entrevistas)	"Eu sinto falta de um link ou de uma explicação para entender como cada um desses fatores. Como esse dado entra? Como ele foi tratado? É positivo ou negativo? É contínuo ou não é contínuo, esse banco né? Ou como é que ele surge? Por que ele pensa a premissa dessa forma? Quais são os dados de conhecimento prévio desse banco? O que faz com que eles valorizem esses dados?"(PO09).

6.1.2 Resultado de explicabilidade da técnica LIME

O presente tópico pretende apresentar os resultados de explicabilidade da técnica LIME. A figura 23 exibe as explicações visuais dos fatores de risco para câncer de colo de útero com a técnica LIME.

Na parte laranja do gráfico são exibidas as características que mais influenciam a previsão de câncer e na parte azul são exibidas as características que menos influenciam a previsão de câncer.

Portanto, como se vê neste exemplo, a mulher tem um alto risco previsto de 0,91. Isto porque os principais fatores para o risco de câncer foram: teste de schiller = sim (1);

Colposcopia = não(0); exame de papanicolau = sim (1); num.parceiro sexuais = 4; idade = 21.

O DST:sífilis = não (0); DSTs: condilomatosa vaginal = não (0); DIU(anos) = 0; contraceptivos hormonais = não (0) foram os principais fatores para o risco de não ter câncer na previsão de 0.09.

É importante destacar que as características do teste de schiller, colposcopia e do exame de papanicolau são binárias (ou seja, "é S ou é N").

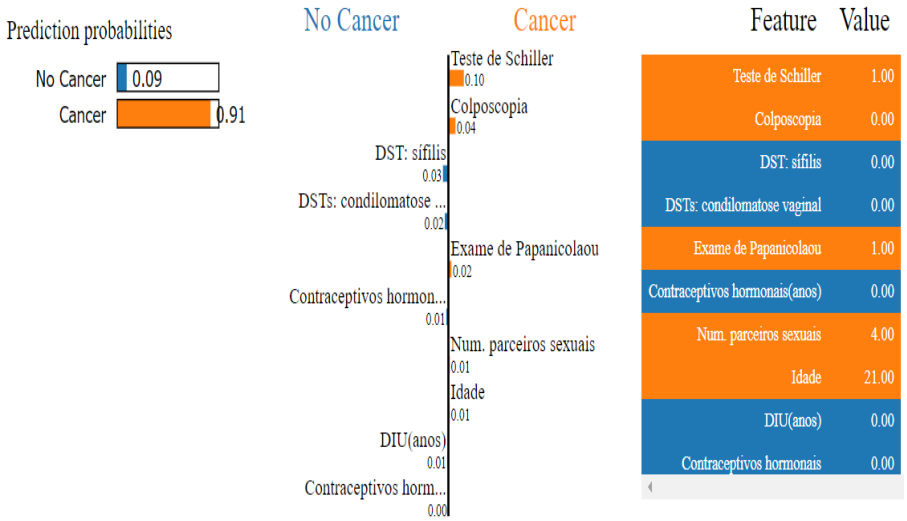

Figura 23 – Técnica LIME. Fatores de risco para câncer de colo do útero

Outrossim, a tabela 10 exibe os resultados da categoria Explicabilidade das Técnicas de IA Explicável para a técnica LIME, como se vê a seguir:

Tabela 10 – Resultados da categoria explicabilidade das técnicas de IA explicável - LIME

Códigos e frequência	***Exemplo de Opinião***
Visual de fácil interpretação (8 entrevistas)	"Gostei mais do LIME. Porque ele é claro e é visualmente completo. (...). Esse aqui num único output, eu tenho a probabilidade da ocorrência de doença, a probabilidade de não ter doença, tenho a lista das variáveis, tenho a distribuição dela influenciando num sentido ou em outro e tenho o peso de cada uma delas. Então eu acho ele mais completo, foi o que mais me interessou."(PO08).
Características com maior influência (12 entrevistas)	"Eu explicaria que o que mais influenciou na predição foram características relacionadas ao exame de Papanicolau, teste de Schiller..."(PO02).

Características com maior/menor influência (6 entrevistados)	"Eu acho que primeiro esse tipo de gráfico que é apresentado aqui, de ter a barrinha do lado direito com câncer com a cor bem clara e não câncer para o lado de cá, foi legal. E aí fica claro quais são os descritores que estão relacionados a não ter câncer nesse caso e quais que foram relacionados a ter câncer também nesse caso e também essa tabela dessa forma dividido em cores. Acho que foi legal."(PO04).
Analogia com a clínica médica (2 entrevistas)	"Eu acho que pelo o que eu entendo, o que ele consegue atribuir atribuir um valor maior ou menor para alguns fatores de risco né, e na presença ou ausência deles né, somar tudo isso e determinar o que a gente faz um pouco de cabeça né. Que é tentar entender o contexto dos exames e entender se a pessoa está com um risco maior ou menor de ter o câncer. E então como proceder em investigações posteriores ou qualquer coisa nesse sentido."(PO12).
Não concordo com o resultado (2 entrevistas)	"Eu entendi o que ele usou para chegar, mas também ainda falta um pouco de sentido em relação a clínica. Porque é a mesma coisa que eu disse anteriormente, ter ou não a Colposcopia, ter ou não o Papanicolau, não seria um resultado que pra mim fazem sentido, para pesar se ele tem câncer realmente. Quer dizer, não ter Papanicolau para mim seria um fator de risco, mas até que no caso dela pode ser que isso tenha algum sentido. Acho que você pegou um caso fora da curva. Acho que você precisa verificar a tendência clínica, até para ajustar a sua coleta de informações."(PO06).
xAI não é explicativa (2 entrevistas)	"Na verdade dizer que o Teste de Schiler, Exame de Papanicolau e etc, não é uma explicação pra mim. Ele não explica o porquê. Então não explica, ele justifica."(PO10).
Técnica não é para leigos (1 entrevista)	"Para mim não tem dificuldade, eu entendo um pouco, trabalho na área, tenho amigos que trabalham na área, fica mais fácil de entender. Mas para um leigo, uma pessoa que nunca viu, eu não sei se seria fácil. O resultado talvez seja fácil, mas a compreensão do que está sendo apresentado aqui, para um leigo, para a população em geral, não sei se fica tão fácil assim de entender."(PO03).

E a seguir segue a tabela 11 que exibe os resultados da categoria Melhoria da Expli-

cabilidade das Técnicas de IA Explicável para a técnica LIME:

Tabela 11 – Resultados da categoria melhoria da explicabilidade das técnicas de IA explicável - LIME

Códigos e frequência	***Exemplo de Opinião***
Detalhar características (6 entrevistas)	"E colocaria qual o tipo de alteração no Papanicolau que surgiu. Por exemplo, doenças de baixo grau e doenças de alto grau."(PO05).
Explicar como interpretar (2 entrevistas)	"E olhando o gráfico e a tabela, eu tive que deduzir a partir da explicação inicial e estou olhando bastante para deduzir e tentando entender."(PO08).
Incluir novos gráficos (1 entrevista)	"Acho que um gráfico ajuda. Quando você coloca as características, com os valores, e um gráfico de barras onde tem os valores no gráfico, acho que isso faz toda diferença. Até para um leigo fica mais fácil de visualizar. Uma apresentação com um gráfico ou com um fluxograma, dependendo do que você quer explicar."(PO03).
Mostrar todas as características (3 entrevistas)	"Assim, coisas a se acrescentar. Acho que, eu não sei se ele usa isso como outras formas de coisas a se ver. Por exemplo, tem outras DSTs, mas HIV é um fator de risco bastante importante para câncer de colo de útero, sei se ele leva em conta isso aí."(PO08).
Incluir peso das características (1 entrevista)	"O peso de repente, porque aqui ele está mostrando os que influenciam mais, mas não o peso". (PO002).
Alterar codificação visual (2 entrevistas)	"Talvez o que faltasse aqui era você ter uma legenda das cores, do laranja e do azul, para mais influência ou menos influência e também de repente até uma gradação dentro do laranja e do azul."(PO04).
Agrupar características pelo peso (1 entrevista)	"...agrupar as caraterísticas pela probabilidade. Para você criar como se fosse uma hierarquia de grupos de características. Se ela apresenta essas características que tenha um peso maior, que essa que um peso intermediário, que essa que tem um peso menor. E quanto isso se relaciona com outras características. Por exemplo, se você juntar duas com um peso menor, daria uma intermediária? Ou duas com um peso menor, daria uma com um peso maior?"(PO01).

Rastreabilidade da informação (3 entrevistas)	"Que tenha acesso no parâmetro para abrir e descobrir como funciona. Ahh, eu quero saber como ele utilizou o Schiller, daí ele diz foi utilizado isso, foi considerado aquilo. Mostrar qual o parâmetro de base utilizado no Schiller. Se eu achei estranho porque ele utilizou isso ou aquilo. Queria poder ver o racional por trás disso, quando eu clicar no nome. Porque ele usou isso daquela maneira. Qual foi o racional por trás daquilo pra dizer se é importante ou não. Eu sinto falta disso."(PO09).

6.1.3 Resultado de explicabilidade da técnica Permutation Importance

Nesta subseção são apresentados os resultados da Explicabilidade da Técnica Permutation Importance.

A figura 24 exibe lista em ordem decrescente das características que mais influenciam na previsão de câncer de colo de útero de acordo com a técnica Permutation Importance.

Como se pode verificar na lista exibida na figura 24, os fatores que mais influenciam na previsão são: Teste de Schiller; Idade; 1º relação; Num. Gestações; Num.parceiros sexuais; exame de papanicolau, dentre outros, totalizando 20 (vinte) fatores (características) para o risco de se ter câncer.

Weight	Feature
0.0577 ± 0.0095	Teste de Schiller
0.0084 ± 0.0108	Idade 1º relação sexual
0.0084 ± 0.0070	Num. gestações
0.0074 ± 0.0095	Num. parceiros sexuais
0.0065 ± 0.0074	Exame de Papanicolaou
0.0056 ± 0.0037	Idade
0.0047 ± 0.0000	DSTs: condilomatose
0.0047 ± 0.0000	Colposcopia
0.0047 ± 0.0059	Contraceptivos hormonais(anos)
0.0047 ± 0.0059	Fuma(anos)
0.0047 ± 0.0000	DSTs
0.0037 ± 0.0037	Contraceptivos hormonais
0.0037 ± 0.0037	DSTs: num de diagnósticos
0.0037 ± 0.0037	DST: sífilis
0.0028 ± 0.0046	Fuma
0.0028 ± 0.0046	DSTs: doença inflamatória pélvica
0.0028 ± 0.0074	DSTs: herpes genital
0.0028 ± 0.0046	DST: condilomatose vulvo-perineal
0.0019 ± 0.0074	DSTs: HPV
0.0009 ± 0.0037	DSTs: condilomatose vaginal
	… *15 more* …

Figura 24 – Técnica Permutation Importance. Fatores de risco para câncer de colo do útero

Assim, passa-se a apresentar a tabela 12 com os resultados da categoria Explicabili-

dade das Técnicas de IA Explicável para a técnica Permutation Importance:

Tabela 12 – Resultados da categoria explicabilidade das técnicas de IA explicável - Permutation Importance

Códigos e frequência	***Exemplo de Opinião***
Visual é de fácil interpretação (2 entrevistas)	"Mas eu achei o Permutation Importance mais claro, porque tem esse tamanho aqui, com os ranges de desvio, e ai quando você não tem um peso, você entende onde estaria. Se fosse por exemplo uma clusterização, você conseguiria onde aquele grupo se encaixaria. E visualmente é mais fácil."(PO02).
Visual é de difícil interpretação (7 entrevistas)	"Esse já é um pouco mais complicado de ser interpretado que o anterior (SHAP). O que eu consigo imaginar é que ele te dá de fato uma ordem de relevância ou de força que cada característica tem para o resultado final. Mas esse peso né, em 0,0195 me diz menos do que o gráfico anterior, do quanto que ele foi relevante para chegar ao resultado final. Então eu entendo que é uma ordem de relevância de cada um deles tem para o resultado final, mas eu acho mais confuso."(PO04).
Características com maior influência (10 entrevistas)	"Ele levou em consideração o Teste de Schiler, foi o que teve o maior peso pra ele chegar a essa predição. Ele também correlacionou a Idade da 1º relação sexual e o número de gestações."(PO07).
Características com maior/menor influência (1 entrevista)	"Então estar mostrando o que pesou mais e o que pesou menos, todos mostraram. Mas eu acho que esse está com um número, não sei se é a forma de mostrar dele, mas o resultado está mais claro."(PO02).
Não concordo com o resultado (5 entrevistas)	"Eu não sei só a Colposcopia, que é uma exame importante, porque ela ficou mais abaixo? Mas talvez porque talvez possa ter algum viés ai nos dados né."(PO11).

xAI não é explicativa (3 entrevistas)	"Que eu entendesse não. Ele me mostrou quais foram os parâmetros que ele utilizou. Mas a explicação do porquê não está aqui. Ele me informa os parâmetros que ele utilizou, mas a relação que ele faz entre o parâmetro e o câncer eu não sei. Isso ele não explicou. Porque olha só, ele me diz aqui Teste de Schiller, idade da 1º relação, número de gestações. Mas percebe que nem o resultado é igual ao outro (LIME). Ele te dar outras informações. Que variável é essa? Eu não sei que conta é essa, se é risco, se é risco relativo, eu não que conta é essa."(PO09).
Técnica não é para leigos (2 entrevistas)	"Como especialista sim, novamente, para especialistas sim. Acho que são todos muito parecidos. Acho que o conhecimento das características."(PO02).

Além disso, apresenta-se a tabela 13 com os resultados da categoria Melhoria da Explicabilidade das Técnicas de IA explicável para a técnica Permutation Importance:

Tabela 13 – Resultados da categoria melhoria da explicabilidade das técnicas de IA explicável - Permutation Importance

Códigos e frequência	***Exemplo de Opinião***
Detalhar características (2 entrevistas)	"Seria bastante interessante ter detalhe de cada exame né."(PO10).
Explicar como interpretar (5 entrevistas)	"Impraticável. Estou tentando entender o output. O peso e a variável. Está em ordem? Está vendo, não consigo nem ver se ele está em ordem. Visualmente ele é muito ruim, a informação está ai, mas eu tenho muita dificuldade de entender. E desvio padrão? Está vendo, eu não consigo nem entender o que é a medida. A primeira medida eu sei que é peso, a segunda medida é desvio padrão ou é faixa? Eu não sei né, não está escrito. A primeira coisa de uma tabela é que ela precisa ser auto explicativa ne, e você não vê isso no output."(PO08).

Incluir novos gráficos (5 entrevistas)	"Eu acho que nesse método, é que apresenta uma espécie de defeito, seria talvez a forma de visualização. A forma gráfica é mais interessante pra gente entender. Aquela coisa de você bater o olho e já ter uma noção boa da informação. Talvez seja isso, esse aqui é necessário olhar com cuidado. Não tem informações percentuais como os outros, que dá pra gente entender bem melhor, o risco ou não. Quanto mais simples pra gente, mais fácil."(PO12).
Incluir peso das características (1 entrevista)	"Apesar de outras pessoas poder achar mais fácil a tabela do que outro, mas eu ainda prefiro o outro. A não ser que você consiga transpor do peso para o percentual de que foi importante para o resultado final. Aí talvez ficaria mais fácil."(PO04).
Alterar codificação visual (1 entrevista)	"O que é o peso, que medidas são essas. Coisas que nos outros modelos que tem uma diagramação visual, fica muito mais fácil de entender."(PO08).
Agrupar características (1 entrevista)	"Que os valores próximos serem agrupados. Ao invés de fatores separados, a partir daqui, fossem feito grupos de fatores, como se colocassem em ordem mesmo. Essas características no grupo 1, essas características da paciente no grupo 2, características no grupo 3. Até para criar fatores de maior peso, fatores intermediários e fatores de menor peso, isso ajudaria bastante o diagnóstico."(PO01).
Incluir nova visualização (1 entrevista)	"Ou num fluxograma, que também ajuda. Toda forma que você consegue melhorar o visual para a pessoa entender melhor o que esta escrito, ao invés de uma tabela. Tabela nunca é didática, não fica claro. Entendeu?"(PO03).
Rastreabilidade da informação (3 entrevistas)	"Mostrar melhor como chegou a esse resultado, porque cada fator desse está nessa ordem. Não fica muito claro, mesmo com o peso, como chegou nesse resultado. Qual a soma das informações que fazem chegar nesse valor aqui. Isso aqui já pé pre estabelecido né?"(PO12).

6.2 Análise e discussão das técnicas de IA explicável

Cumpre analisar como as técnicas LIME, SHAP e Permutation Importance apresentam suas explicações.

Em primeiro lugar, aplicou-se a técnica SHAP. Importa dizer que no caso em estudo foi utilizado um algoritmo de florestas aleatórias (random forests), então foi usado a variação pertinente a modelos de aprendizado de máquina baseados em árvores, qual seja, o TreeSHAP [27]. O SHAP apresenta as características que mais influenciam ou menos influenciam num gráfico de barra vertical, de acordo com a previsão do modelo.

Portanto, de acordo com a técnica utilizada, quanto maior a precisão do modelo, mais características influentes positivamente são exibidas ou vice versa. Neste ponto, remetendo-se à imagem da figura 22, encontra-se um modelo com previsão de 0.91, sendo certo que essa previsão é local, ou seja, a previsão se aplica para instância de dados específicos (paciente).

No exemplo da figura 22, encontra-se claramente 07 (sete) características aproximadamente, sendo 05 (cinco) características que mais influenciam (quais sejam: Teste de Schiler, Exame de Papanicolau, Num.parceiros sexuais, Idade e Idade da 1º relação sexual) e 02 (duas) que menos influenciam (quais sejam: Hinselmann e Num.gestações).

Em seguida, passa-se à análise da técnica LIME e, para isso, faz-se necessário verificar a imagem da figura 23, onde existem 02 (duas) formas diferentes de apresentação: a) tabela e b) barra vertical. A previsão para o câncer é de 0.91 e para sua não ocorrência de câncer é de 0.09.

Neste caso, a previsão também é local, já que se aplica para uma única instância (paciente). Com esta a técnica, verifica-se de forma clara 10 (dez) características, sendo 05 (cinco) que mais influenciam (quais sejam: Teste de Schiller, Colposcopia, Exame de Papanicolau, Num.parceiros sexuais e Idade) e 05 (cinco) que menos influenciaram (quais sejam: DST: sífilis, DSTs: condilomatose vaginal, Contraceptivos Hormonais (anos), DIU (anos) e Contraceptivos Hormonais).

Por último, com a técnica Permutation Importance, a explicação é apresentada por meio de tabela, onde as características mais influentes para a decisão são exibidas em seu topo, de modo que as causas são listadas da mais influente para a menos influente. Com esta técnica, o valor do score não é exibido e o modelo é global, pois se considera todo o conjunto de dados para a emissão do resultado.

A técnica do Permutation Importance apresenta o peso (weight), que considera a quantidade de aleatoriedade calculada, sendo o número após o ± o medidor do desempenho variante de uma reorganização para a outra. No exemplo trazido pela imagem da figura 24, constata-se aproximadamente 20 (vinte) características (quais sejam: Teste de Schiller, Idade 1º relação sexual, Num. gestações e etc) em ordem de importância na explicação para a previsão do modelo.

A tabela 14 a seguir exibe os resultados da categoria Explicabilidade das Técnicas de IA Explicável de maneira consolidada:

Tabela 14 – Resultados da categoria explicabilidade das técnicas de IA explicável

Código	***SHAP***	***LIME***	***PI***	***Total***
Visual de fácil interpretação	9 (PO01, PO03, PO04, PO05, PO06, PO07, PO09, PO11, PO12)	8 (PO03, PO04, PO05, PO07, PO08, PO09, PO11, PO12)	1 (PO02)	10
Visual é de difícil interpretação	0	0	7 (PO04, PO05, PO06, PO07, PO08, PO09, PO12)	7
Características com maior influência	12 (todos)	12 (todos)	10 (PO01, PO02, PO03, PO04, PO06, PO07, PO09, PO10, PO11, PO12)	12
Características com menor influência	5 (PO01, PO02, PO05, PO07, PO09)	6 (PO02, PO04, PO06, PO07, PO09, PO12)	1 (PO02)	8
Analogia com a clínica médica	0	2 (PO08, PO12)	0	2
Não concordo com o resultado	1 (PO06)	2 (PO06, PO09)	5 (PO01, PO05, PO06, PO07, PO11)	8
xAI não é explicativa	2 (PO09, PO10)	2 (PO09, PO10)	3 (PO05, PO09, PO10)	3

Figura 25 – Categoria Explicabilidade das técnicas de IA Explicável

Técnica não é para leigos	2 (PO02, PO10)	1 (PO03)	2(PO02, PO03)	3

Dos resultados mais gerais obtidos em pesquisa com a utilização da Teoria fundamentada em dados, conforme exibido pela figura 25, pode-se relacionar que são basicamente 03 (três) os motivos que levam os usuários a terem uma melhor percepção de explicabilidade das técnicas de IA explicável: a) **visual de fácil interpretação**; b) **visualizar as características com maior/menor influência** e c) **analogia com a clínica médica**.

Além disso, há 04 (quatro) motivos para se relacionar à percepção de pouca explicabilidade das técnicas de IA explicável: a) visual é de difícil interpretação, Não concordo com o resultado; b) xAI não é explicativa e c) técnica não é para leigos.

Dentre os motivos que levam a uma melhor percepção de explicabilidade, o primeiro deles está em **visual de fácil interpretação**, já que a visualização da explicação permite ao usuário interpretar os fatores de influência de forma simples e rápida, sem a necessidade de prévio treinamento da técnica de explicabilidade, além do conhecimento do domínio. Essa motivação foi encontrada com maior intensidade ao utilizar as técnicas SHAP e LIME, donde se permite concluir que essas técnicas são simples de interpretar.

Em contrapartida, a visualização das **características com maior/menor influência** é um acréscimo, já que a maneira mais usual dos usuários interpretarem os resultados nas técnicas estudadas são pela listagem das **características com maior influência**, ou seja,

a listagem das características que mais influenciam a classificação.

Quando analisadas as características com maior/menor influência, é possível entender qual é a listagem das características que mais influenciam a classificação, bem como as características que menos influenciam a classificação, ou seja, numa única visualização é possível visualizar os fatores à favor e contra.

Essa motivação está mais presente nas técnicas LIME e SHAP, o que sugere uma maior percepção de explicabilidade por meio dessas técnicas.

Em **analogia à clínica médica**, eis a motivação que relaciona a explicação da técnica com o raciocínio que o profissional usa na prática clínica. Essa motivação só foi citada para a técnica LIME, o que pode sugerir um maior grau de explicação por essa técnica, uma vez que oferece mais elementos à tomada de decisão pelos especialistas no domínio.

Dentre os motivos que trazem uma percepção de pouca explicabilidade das técnicas de IA explicável está o código **visual é de difícil interpretação**, exclusivamente citado na técnica Permutation Importance. A dificuldade apontada pode estar relacionada à medida peso (weight), já que pouco entendida pelos usuários.

Além disso, o código **Não concordo com o resultado** foi citado em todas as técnicas, sendo uma crítica mais intensa para a técnica Permutation Importance. Essa percepção pode estar relacionada ao conjunto de dados usados, que possui limitações quanto ao seu detalhamento e, no caso específico do Permutation Importance, por se tratar de um resultado global, a explicação baseia-se em todo o conjunto de dados e não a uma previsão específica.

Entre os motivos de pouca explicabilidade, constata-se que o mais incisivo está relacionado à percepção de que o **xAI não é explicativa**. Embora essa opinião esteja presente em todas as técnicas, é no Permutation Importance que aparece em maior grau.

Isto pode decorrer da opinião de que essas técnicas só exibem os fatores que mais influenciam e não mostram os mecanismos internos de decisão, já que não existe rastreabilidade entre o conjunto de dados, o processamento e a saída da explicação da técnica, ou seja, o "porquê"não é mostrado como explicação para os usuários.

Como último motivo de pouca explicabilidade, tem-se a opinião de que essas **técnicas não são para leigos**, ou seja, ela é mais indicada para um especialista no domínio. Isto porque essa opinião é distribuída de maneira igualitária por todas as técnicas e pode estar relacionada ao domínio discutido, que é considerado um domínio crítico e pouco acessível aos usuários leigos.

Por outro lado, a tabela 15 a seguir exibe os resultados da categoria Melhoria da Explicabilidade das Técnicas de IA Explicável de maneira consolidada:

Tabela 15 – Resultados da categoria melhoria da explicabilidade das técnicas de IA explicável

Código	***SHAP***	***LIME***	***PI***	***Total***
Detalhar características	5 (PO05, PO06, PO07, PO10, PO12)	6 (PO05, PO06, PO07, PO08, PO09, PO10)	2 (PO07, PO10)	7
Explicar como interpretar	2 (PO03, PO09)	3 (PO04, PO08, PO09)	5 (PO04, PO05, PO06, PO08, PO09)	6
Incluir novos gráficos	0	1 (PO03)	5 (PO03, PO06, PO08, PO11, PO12)	5
Mostrar todas as características	2 (PO04, PO08)	3 (PO05, PO11, PO12)	0	5
Incluir peso das características	2 (PO02, PO04)	1 (PO02)	1 (PO04)	2
Alterar codificação visual	0	2 (PO04, PO06)	1 (PO08)	3
Agrupar características	1 (PO06)	1 (PO01)	1 (PO06)	2
Incluir nova visualização	0	0	1 (PO03)	1
Incluir rastreabilidade da informação	3 (PO09, PO10, PO12)	3 (PO09, PO10, PO12)	3 (PO09, PO10, PO12)	3

Dos resultados obtidos em pesquisa com a utilização da Teoria fundamentada em dados, encontra-se 08 (oito) sugestões de melhoria das técnicas de IA explicável com potencial de melhoria nas explicações providas pelas técnicas de IA explicável, quais são elas: a) Detalhar características, b) explicar como interpretar, c) incluir novos gráficos, d) mostrar todas as características, e) incluir peso das características, f) alterar codificação visual, g) agrupar características, h) incluir nova visualização e i) incluir rastreabilidade da informação.

A sugestão de **detalhar características** deve estar relacionada ao fato de que algumas características importantes para o domínio do câncer de colo de útero (Colposcopia, Teste de Schiller e Exame de Papanicolau) são muito usadas como fator de decisão na clínica médica, em que pese serem características binárias (ou seja, "é S ou é N").

A utilização de características binárias se mostra um limitador oriundo do conjunto de dados utilizados para a construção do modelo. Como se verifica, o conjunto de dados normalmente utilizado não possui características ou atributos adicionais que permitam detalhar a colposcopia, o teste de schiller e o exame de papanicolau e essa ausência de especificidade pode estar intimamente ligada à percepção de **não concordância com a explicação**, amplamente citado em todas as técnicas.

O resultado **Explicar como interpretar** é citado em todas as técnicas, sendo muito mais perceptível no Permutation Importance, sendo possível afirmar que se trata de uma necessidade real, já que em nenhuma das técnicas estudadas existe algum tipo de ajuda ou informação de "como interpretar", ficando tal tarefa a cargo da intuição dos usuários.

Portanto, essa sugestão pode estar relacionada à percepção de que o **visual é de difícil interpretação**, principalmente no Permutation Importance, uma vez que os usuários não conseguem entender a métrica peso (weight) apresentada na explicação.

As melhorias **incluir novos gráficos**, **alterar codificação visual** e **incluir nova forma de visualização** são sugestões relacionadas à visualização da informação e são citadas em maior grau ao utilizar a técnica Permutation Importance, o que pode significar que a saída da técnica oferece poucos elementos de explicabilidade e o **visual é de difícil interpretação**.

Outras melhorias, como **agrupar características pelo peso**, **incluir peso das características** e **incluir rastreabilidade da informação** podem estar fortemente relacionadas à percepção de que o visual é de difícil interpretação, Não concordo com o resultado e principalmente à percepção de pouca explicabilidade das técnicas xAI, já que essas sugestões estão relacionadas ao anseio dos usuários em conhecer os mecanismos algorítmicos das técnicas.

6.3 Confiança na Inteligência Artificial

Nesta seção estão descritos os resultados da percepção de confiança dos usuários especialistas no domínio médico em relação à inteligência artificial.

A tabela 16 exibe os resultados da categoria confiança em inteligência artificial:

Tabela 16 – Resultados da categoria confiança em inteligência artificial

Códigos e frequência	***Exemplo de Opinião***
#01-Testar/validar (5 entrevistas)	"ter um range grande de pessoas para pode testar cada vez mais o programa para ver se chega a algum consenso ali para poder liberar né."(PO05).
#02-Considerações éticas (1 entrevista)	"Na medicina vai envolver muita coisa relacionado a IA e ML. Nós confiamos, mas a princípio é testar e fazer estudos de como implantar os métodos. Mas eu tendo a confiar e acreditar. Eu confio nos resultados gerados. Mas não confio na ética."(PO10).
#03-Curadoria de dados (4 entrevistas)	"O principal é você saber selecionar todos os dados né. Porque se você não tiver os dados certos, você pode acabar tendo um viés que acaba interferindo na Inteligência Artificial. Acho que a IA pode ajudar em muita coisa, mas eu acho que a questão é saber selecionar os dados e saber se aqueles dados são confiáveis né. Acho que isso é o mais importante. Não a IA em si, mas quem selecionou os dados para ser utilizado na IA."(PO11).
#04-Maior assertividade (1 entrevista)	"Confio. Por que eu confio? Porque a taxa de assertividade ela é alta e vem sendo provado ao longo do tempo."(PO07).
#05-Não substitui o profissional de saúde (4 entrevistas)	"Então, eu não sou contra. Mas eu acho que é uma coisa que deve ser aperfeiçoado daqui pra frente e eu acho que não substitui a experiencia de um profissional."(PO05).
#06-Uso IA como apoio (2 entrevistas)	"Eu acho que pode ser uma coisa que ajude numa triagem bem inicial. Como por exemplo a análise da Citologia hoje, pode ser feito por técnicos e quando o técnico vê algo diferente ele chama o profissional médico para ele ver o diagnóstico. Eu acho que isso aqui pode ser tipo uma peneira mesmo dos casos mais prováveis e daí o médico confirmar se é isso ou não. E a partir daí passar a conduta."(PO05).
#07-Viés algorítmico (3 entrevistas)	"O principal é você saber selecionar todos os dados né. Porque se você não tiver os dados certos, você pode acabar tendo um viés que acaba interferindo na Inteligência Artificial."(PO11).

#08-Conhecer limitadores da tecnologia (1 entrevista)	"Se você não tiver um parâmetro muito claro para entender, o porquê e onde a máquina pode errar, você vai se lascar. A maioria das máquinas que eu uso, eu sei os fatores críticos. Então o fator crítico eu tenho que olhar. O fator crítico é o limitador. Eu também tenho fator crítico, o ser humano tem essas questões. Quando eu chego no meu limite eu tenho que checar e rechecar. Então quando eu vou rechecar a máquina? Ou vou aceitar tudo passivamente? Eu sinto falta dessas questões."(PO09).
#09-Treinar profissionais (2 entrevistas)	"É viável. Só acho que tem que melhorar bastante e não só isso né, é necessário treinar os profissionais e ter um range grande de pessoas para pode testar cada vez mais o programa para ver se chega a algum consenso ali para poder liberar né."(PO05).
#10-Big data é facilitador (3 entrevistas)	"Porque acho que os dados, um número grande de dados eles tem uma representatividade, nos dá uma informação que talvez a gente não consiga visualizar sozinho. Acho que isoladamente não é suficiente, não basta. Mas eu acho que eles têm muito valor."(PO06).
#11-Necessidade do fator humano (2 entrevistas)	"Então assim. Tem coisas que eu acho que não vai substituir. O relacionamento humano nunca vai ser substituído pela máquina. A gente vê o problema do excesso do uso de máquina."(PO09).

Já a tabela 17 exibe os resultados da categoria confiança em inteligência artificial por entrevistado:

Tabela 17 – Resultados da categoria confiança em inteligência artificial por entrevistado

Código	***Entrevistados***	***Total***
#01-Testar/validar	5 (PO05, PO07, PO09, PO10, PO12)	5
#02-Considerações éticas	1 (PO10)	1
#03-Curadoria de dados	4 (PO06, PO09, PO11, PO12)	4
#04-Maior assertividade	1 (PO07)	1
#05-Não substitui o profissional de saúde	4 (PO05, PO09, PO11, PO12)	4

#06-Uso IA como apoio	2 (PO01, PO05)	2
#07-Viés algorítmico	3 (PO06, PO09, PO11)	3
#08-Conhecer limitadores da tecnologia	1 (PO09)	1
#09-Treinar profissionais	1 (PO05)	1
#10-Big Data é facilitador	3 (PO04, PO06, PO09)	3
#11-Necessidade Fator Humano	2 (PO09, PO12)	2

O objetivo desta pesquisa é explicar previsões, recomendações e outras saídas de IA para os seus usuários a fim de proporcionar confiança no resultado da pesquisa.

Portanto, a forma como se oferece explicações sobre o funcionamento interno do seu sistema de Inteligência Artificial (IA) pode influenciar profundamente a experiência do usuário com seu sistema e sua utilidade na tomada de decisões.

Embora a IA seja uma importante ferramenta de suporte e apoio para o médico tomar decisões, os entrevistados confiam na inteligência artificial, mas com algumas ressalvas.

De acordo com a pesquisa, o caminho da inteligência artificial na medicina deve se tonar uma ferramenta que auxilia os profissionais a executarem seu trabalho com **maior assertividade** e rapidez, mas jamais substituindo as relações interpessoais. A experiência com saúde não pode ser totalmente robotizada e automatizada, ou seja, o **fator humano** ainda é fundamental. Empatia, por exemplo, vital no relacionamento médico-paciente, é atributo humano.

De acordo com a entrevista PO09: "(...). Então assim o fator humano é, nem tudo a pessoa vai responder de verdade para a máquina, pode omitir mais ou menos, depende do quanto ela confia na questão da segurança dos dados dela. Tem coisas que você conta para um médico e não conta pra outro. O médico pode não valorizar outro vai valorizar, então vai depender do seu background."(PO09).

Desta forma, o principal objetivo do desenvolvimento de aplicações com inteligência artificial é aprimorar e auxiliar atividades humanas, mas de acordo com os profissionais entrevistados, a IA **não substitui o trabalho do profissional de saúde**, pois, afinal, um

diagnóstico não é baseado apenas em aprendizagem de máquina, por exemplo.

Todo médico, seja qual for a sua especialidade, avalia diferentes aspectos para chegar a um resultado, como o histórico do paciente, sintomas e sinais, patologias existentes, predisposições genéticas e familiares e, até, questões psicológicas, sociais e ambientais.

Assim, como é possível aumentar a confiança na inteligência artificial? Segundo os participantes da pesquisa, os principais fatores relatados são: a) **testar/validar**, b) **curadoria de dados**, c) **conhecer limitadores da tecnologia** e d) **treinar profissionais**.

Desta maneira, ao instituir as ações elencadas pelos participantes da pesquisa será possível evitar o **viés algorítmico** e os problemas decorrentes da **falta de considerações éticas** que tanto afetam a confiança dos usuários na inteligência artificial.

E, claro, a medida que mais IA for utilizada no âmbito da saúde, a confiança nesse modelo ganhará mais força, até porque estamos lidando com pacientes e vidas reais.

6.4 Recomendação para a implantação de técnicas de IA explicável no domínio médico

Nesta seção estão descritos os resultados das recomendações para a implantação de técnicas de IA explicável no domínio médico.

A tabela 18 contem os resultados da pesquisa que justificam a recomendação para a implantação de técnicas de IA explicável no domínio médico:

Tabela 18 – Resultados da categoria recomendação para a implantação de técnicas de IA explicável no domínio médico

Códigos e frequência	***Exemplo de Opinião***
#01 - Popularizar as técnicas de IA explicável (2 entrevistas)	"Eu acho que a sua popularização. Não só que eles tenham sua acurácia medida e divulgada, que eles sejam popularizados para que a gente possa utiliza-los na pratica clínica. Seja por exemplo, o próprio usuário do sistema de saúde avaliando pelos seus sinais e sintomas, se precisa procurar o médico ou não. Seja por um setor de triagem que precisa se um paciente precisa de atendimento mais rápido ou não num sistema de agendamento do que outro."(PO08).

#02 - Participação do especialista do domínio (1 entrevista)	"Eu acho que tem que ter um conhecimento clínico para escolher os fatores. Sem conhecimento clínico dos fatores é muito difícil desenhar, partindo dos dados que já tem. Eu acho que é importante pensar na a fisiopatologia da doença para você conseguir buscar quais fatores que vão te ajudar nisso. Pelo menos no primeiro momento. Porque o que vi aí de crítica é que tem elementos e variáveis que para mim não fazem tanto sentido clínico, para justificar um risco de câncer."(PO06).
#03 - Selecionar, usar e testar diversos conjunto de dados (3 entrevistas)	"Eu acho que de repente, você disse que usou dados de um hospital de Caracas né? Eu acho de repente você poderia testar com dados de outras unidades né, trabalhar com dados reais da população brasileira, ou até mesmo de outros países. Quanto mais troca você tiver, menor será o risco né. Claro que existem riscos que são comuns a todas as populações né, mas tem riscos muito pertinentes a cada país. Por exemplo na África tem muito HIV positivo. Ali o maior fator de risco é ser imunossuprimido. Outros são países com alto nível de tabagismo ou menor acesso a saúde para prevenção de câncer, como exame de Papanicolau. Tem pais que tem maior cobertura vacinal para HPV, tem pais que tem menor cobertura. As particularidades é que temos que mudar né. Acho que é mais avaliar de uma forma geral e da a melhor forma de montar o programa de uma forma igualitária. Acho que é isso."(PO05).
#04-Evidência da confiança e aplicabilidade do modelo (5 entrevistas)	"(...) eu acho que primeiro tem que ir fazer a validação de estudos para outros tipos de câncer, outros tipos de bancos de dados, mais complexos inclusive. E aí você sempre tendo esse valor alto de predisposição, você convence. Então eu acho que é o normal, você tá indo por um caminho, tá fazendo com um tipo, um banco de dados. Acho que teria que validar em outra corte, em outro tipo de amostra né, de paciente. Para esse tipo tumoral e depois expandir para outros tipos tumorais."(PO04).

Já a tabela 19 exibe os resultados da categoria recomendações para a implantação de técnicas de IA explicável no domínio médico por entrevistado:

Tabela 19 – Resultados da categoria recomendações para a implantação de técnicas de IA explicável no domínio médico por entrevistado

Código	***Entrevistados***	***Total***
#01 - Popularizar as técnicas de IA explicável	2 (PO03, PO08)	2
#02 - Participação do especialista do domínio	1 (PO05)	1
#03 - Selecionar, usar e testar diversos conjunto de dados	3 (PO04, PO05, PO09)	3
#04-Evidência da confiança e aplicabilidade do modelo	5 (PO01, PO04, PO07, PO09, PO11)	5

Como se constata na tabela 18 acima, os principais fatores relatados pelos participantes e que justificam a implantação de IA explicável no domínio médico são:

a) Popularizar as técnicas de IA explicável;

b) Participação do especialista do domínio;

c) Selecionar, usar e testar diversos conjunto de dados;

d) Evidência da confiança e aplicabilidade do modelo.

Nesta pesquisa foi verificado que os usuários possuem pouco conhecimento acerca da existência de técnicas de explicação, de maneira que a **popularização** dessas técnicas foi apontada como um fator primordial na medicina. Neste passo, sugere-se que, além da acurácia medida e divulgada, as técnicas sejam popularizadas para que profissionais do domínio médico possam utilizá-los na prática clínica. Isto porque, embora a IA faça parte de nossas vidas nos mais diversos domínios, esses conceitos ainda são poucos conhecidos por pessoas não especializadas na área.

A **participação do especialista do domínio** foi apontada como fator de grande importância em todo o processo de solução, ou seja, desde a formulação do problema, hipóteses de resolução até a análise de resultados. Portanto, é sempre necessário incluir um profissional que tenha conhecimento clínico e de fisiopatologia nos processos orgânicos

da doença para se perquirir quais são as características importantes. A ausência do conhecimento clínico sobre os fatores de risco da doença torna muito difícil o desenho e a modelagem a partir dos dados existentes, assim como identificar o viés nos dados. **Selecionar, usar e testar diversos conjuntos de dados**. Os dados que treinam os algoritmos de aprendizagem de máquina são criados, limpos, rotulados e anotados por humanos, assim como a construção dos modelos. Assim, testar e usar dados de várias origens e geografias, com dados reais da população brasileira ou, até mesmo de outros países, nos dá mais chance de montar um sistema igualitário e com menor risco de viés.

Embora existam riscos comuns a todas as populações, alguns riscos são particulares de cada localidade/País. Ora, particularidades como alto nível de tabagismo, maior ou menor acesso à programas de saúde preventivos do câncer (como o exame de Papanicolau) ou, ainda, maior ou menor cobertura vacinal para HPV são variantes que interferem no resultado e, portanto, devem ser levadas em consideração.

A título exemplificativo, temos o continente africano onde há muitos pacientes portadores de HIV positivo, portanto, tal especificidade resulta em pacientes com maior fator de risco de serem imunossuprimidos.

Evidência da confiança e aplicabilidade do modelo se faz necessário, pois embora a IA venha sendo utilizada na medicina para aprimorar o diagnóstico, prognóstico e tratamento, ela ainda gera muita desconfiança, o que descredencia as informações obtidas por meio do aprendizado de máquinas.

7. Principais Achados e Recomendações para o Melhoria das Técnicas de IA Explicável

Este capítulo apresenta os principais achados e recomendações para o desenvolvimento de sistemas de IA explicável e sugestões de implantação desses métodos no domínio médico, elaboradas a partir de dados obtidos na interpretação e análise dos dados no capítulo anterior.

7.1 Principais achados das técnicas de inteligência artificial explicável

Esta seção exibe os achados obtidos em pesquisa com a utilização da Teoria fundamentada em dados quanto à percepção de explicabilidade das técnicas de IA explicável.

1 - Técnicas xAI agnósticas não explicam.

A maioria das explicações das técnicas de IA explicável agnóstica é limitada a uma lista ou representação gráfica das principais características que influenciam uma decisão e a sua importância relativa, o que é chamado de importância de características ou de principais fatores de contribuição.

Especialistas no domínio sugerem que essa maneira de apresentação dos resultados das técnicas de IA Explicável agnósticas oferece somente uma justificação e não uma explicação.

A ocorrência do código ‘xAI não explica’ nas entrevistas é recorrente como no exemplo a seguir: "Não. Que eu entendesse não. Ele me mostrou quais foram os parâmetros que ele utilizou. Mas a explicação do porquê não está aqui. Ele me informa os parâmetros que ele utilizou, mas a relação que ele faz entre o parâmetro e o câncer eu não sei. Isso ele não explicou. E provavelmente todos os métodos terão o mesmo defeito. Eles todos vão me dizer qual que tem influência positiva ou negativa, e todos vão me falhar em explicar

o porquê. Então ele não explica, só informa."(PO09).

O importante achado na pesquisa é que a apresentação da importância das características pelas técnicas de IA explicável transmite pouco a título de explicação, embora possa ser informativa.

Desta maneira, certos tipos de problemas, principalmente em domínios críticos, não podem ser facilmente entendidos pela quantificação e listagem de alguns fatores. Isto porque a apresentação de importância das características ignora amplamente os detalhes de interações entre características, portanto, mesmo as explicações mais ricas baseadas nessa abordagem são limitadas a relativamente simples afirmações, ou seja, o "porquê"não é respondido.

Essa questão levantada nas entrevistas vai no mesmo sentido da definição de Biram e Cotton (2017) [10] em que uma justificativa explica por que uma decisão é boa, mas não necessariamente visa dar uma explicação do processo real de tomada de decisão.

Ainda de acordo com eles, explicabilidade é algo fortemente relacionado à noção de interpretabilidade, pois um sistema interpretável seria aquele cujos resultados são compreensíveis para nós humanos, seja por meio da inspeção do sistema, seja por meio de alguma explicação produzida durante o seu funcionamento.

Além disso, eles estabelecem uma distinção entre interpretabilidade e a noção de justificação, cujo objetivo seria explicar porque a decisão tomada pelo sistema pode ser aceita como uma boa decisão. Ou seja, justificabilidade e interpretabilidade seriam capacidades complementares.

2 - Mostrar na técnica de explicação as características de maior/menor influência na explicação pode melhorar a explicabilidade.

A maneira comum de apresentação das explicações numa técnica de explicação independente de modelo (agnóstica) é listando as características pela ordem de importância (Feature Importance), onde são exibidos: (a) características com evidências positivas, que seriam as características que mais influenciam a classificação e (b) características com evidências negativas, que seriam as características que menos influenciam a classificação.

Essas maneiras de interpretar os resultados das técnicas de inteligência artificial explicável (SHAP, LIME e Permutation Importance) foram bem representadas nas entrevistas, através do código ‘características com maior influência’ e ‘características maior/menor influência’.

A ocorrência do código ‘características com maior influência’ significa que os especialistas no domínio interpretam o resultado (em forma de gráfico), citando as características com evidências positivas para o câncer.

A ocorrência do código ‘características com maior/menor influência’ significa que os especialistas no domínio interpretam o resultado (em forma de gráfico), citando as características com evidências positivas e negativas para o câncer.

Alguns trabalhos recentes da academia sugerem que a explicabilidade das técnicas de IA explicável agnósticas ainda é algo distante, uma vez que essas técnicas post-hoc não têm acesso a nenhuma propriedade interna do modelo, como pesos, restrições ou suposições [18]. Portanto, é possível melhorar a explicabilidade oferecendo elementos que ajudem o usuário a melhorar sua percepção de explicabilidade.

Um achado na pesquisa é a possibilidade de ter uma melhor compreensão dos resultados nas técnicas de IA explicável quando se tem uma visualização integrada das características que mais influenciam e das características que menos influenciam uma decisão numa única tela, ou seja, exibindo ‘características com maior/menor influência’.

Neste sentido, seguem alguns relatos: "Mas não aparecem todas as variáveis, só aparecem parte delas, então o outro (LIME) tem a vantagem de mostrar todas as variáveis que influenciaram ou deixaram de influenciar na predição. Então me parece que o outro (LIME) me dar mais transparência, explica melhor."(PO08).

Na técnica de IA explicável LIME, o resultado (gráfico) é apresentado com as características de maior/menor influência na explicação, permitindo que os usuários visualizem as características de maior risco e menor risco numa única visualização, aumentando assim o grau de explicabilidade.

Em compensação, na técnica SHAP, quanto maior a precisão do modelo, menos características que influenciam negativamente na predição são mostradas no gráfico.

Já no Permutation Importance, são exibidos de maneira clara, no máximo, 15 (quinze) características em ordem crescente de maior influência.

3 - Técnicas de IA explicável agnósticas globais são menos explicáveis

Explicações globais visam fornecer uma apresentação mais holística de como o sistema funciona para todo o conjunto de dados ou para coleções de instâncias.

No entanto, as explicações globais, geralmente, não são interpretáveis ou simplistas demais para representar o modelo original [6].

O Permutation Importance é o único representante de técnica agnóstica global, ou seja, fornece uma visão global altamente compactada do comportamento do modelo.

Com a falta de previsibilidade específica (uma única instância) - já que o Permutation Importance oferece uma explicação global - ocorre uma dificuldade de interpretação e compreensão da técnica Permutation Importance e essa dificuldade foi bem evidenciada nas entrevistas, dada as ocorrências de códigos relacionados a pouca explicabilidade e interpretabilidade, tais como: 'visual de difícil interpretação', 'não concordo com a explicação', 'xAI não explica' e 'técnica não é para leigos'.

A ocorrência do código 'visual difícil de interpretação' significa que os especialistas no domínio têm dificuldade de compreender o resultado (em forma de gráfico), de acordo com o relato dos entrevistados: "Impraticável. Estou tentando entender o output. O peso e a variável. Está em ordem? Está vendo, não consigo nem ver se ele está em ordem. Visualmente ele é muito ruim, a informação está ai, mas eu tenho muita dificuldade de entender. E desvio padrão? Está vendo, eu não consigo nem entender o que é a medida. A primeira medida eu sei que é peso, a segunda medida é desvio padrão ou é faixa? Eu não sei né, não está escrito. A primeira coisa de uma tabela é que ela precisa ser auto explicativa né, e você não vê isso no output."(PO08).

A ocorrência do código 'não concordo com a explicação' significa que os especialistas não concordam com o resultado, pois não é compatível com a prática clínica desses especialistas.

Já a ocorrência do código 'XAI não explica' significa que a técnica oferece pouca ou nenhuma explicação.

E a ocorrência de 'técnica não é para leigos', como o próprio nome diz, significa que a técnica Permutation Importance não é adequada para usuários leigos, mas somente especialistas no domínio.

Nesta pesquisa, chega-se à conclusão de que as técnicas de IA explicável agnósticas globais oferece pouco grau de explicabilidade, uma vez que a explicação baseada em todo o conjunto de dados oferece pouco insight para uma decisão individual. E, para explicabilidade humana, por exemplo, explicações locais podem ser usadas para explicar a conexão entre uma única instância de entrada e a saída da máquina resultante.

4 - Incluir mecanismos de rastreabilidade na técnica de explicação pode melhorar a explicabilidade

Doran e colegas [61] classificam os sistemas de inteligência artificial explicável em

03 (três) tipos:

1) Sistemas Opacos: são aqueles que não oferecem informações sobre seus mecanismos algorítmicos;

2) Sistemas interpretáveis: são aqueles que os usuários podem analisar matematicamente seus mecanismos algorítmicos;

3) Sistemas compreensíveis: são aqueles que emitem símbolos, permitindo explicações orientadas pelo usuário sobre como chegar a uma conclusão, ou seja, compreender por que uma certa saída está associada a uma certa entrada.

De acordo com esta classificação, as atuais técnicas de IA explicável agnóstica se classificam como sistemas opacos, ou seja, não oferecem informações sobre seus mecanismos algorítmicos [10], pois, encontra-se elementos na pesquisa através da ocorrência dos seguintes códigos de sugestão: 'Explicar como interpretar', 'incluir peso de contribuição das características', 'agrupar características por peso' e 'rastreabilidade da informação'.

É bem verdade que não seja possível construir sistemas de IA explicável para atender aos requisitos de oferta de informações sobre seus mecanismos algorítmicos – já que essas técnicas não possuem acesso aos mecanismos internos dos modelos como, por exemplo, peso e estrutura-, mas é possível incluir elementos ou símbolos que tem potencialidade de melhoria na compreensão dos resultados nas técnicas de inteligência artificial explicável.

5 - Mostrar na técnica de explicação todas as características da predição pode melhorar a compreensibilidade.

Como dito anteriormente, a maneira comum de se apresentar os fundamentos de uma técnica de explicação independente de modelo (agnóstica) se dá através do elenco de suas características listadas pelo seu grau de importância ou pelos seus principais fatores determinantes.

A fórmula citada acima está presente nas técnicas SHAP, LIME e Permutation Importance, pois elas não revelam todas as características usadas na predição do modelo, ou seja, não apontam o fator determinante para ajudar o especialista do domínio à obtenção de um melhor entendimento dos resultados.

De acordo com as informações obtidas na pesquisa, se fosse possível a visualização de todas as características usadas na predição, o resultado da técnica de IA explicável levaria a elementos primordiais para a melhor compreensão dos resultados, pois, desta forma, poderia ser analisado o motivo pelo qual certas características não são influentes

e, por consequência, associá-las (fazer analogia) à prática clínica.

Tal conclusão ficou evidenciada com a alta ocorrência do código de melhoria 'mostrar características', denotando a necessidade de se exibir todas as características usadas na predição.

Contudo, a limitação apontada está presente nas 03 (três) técnicas usadas na pesquisa, já que:

• O SHAP apresenta em seu resultado (gráfico de saída) no máximo 10 (dez) características.

• Já o LIME apresenta aproximadamente 10 (dez) características.

• E, por fim, o Permutation Importance apresenta no máximo 20 (vinte) características em ordem crescente de maior influência.

Conclui-se, pois, que de acordo com os resultados das entrevistas, o melhor entendimento dos resultados obtidos por meio do uso de técnicas de inteligência artificial explicável ocorrerá quando for possível visualizar todas as características usadas na predição do modelo em uma única tela.

6 - A melhoria da visualização da técnica de explicação pode melhorar a explicabilidade

Atualmente, a visualização de dados é feita por meio gráfico, porém o objetivo é simplificar a visualização desses dados para, então, promover a sua compreensão, além de transmitir conceitos e idéias.

Entende-se que a visualização de dados é boa quando relacionada a uma boa legibilidade de informações.

Contudo, todas as técnicas de IA explicável agnóstica utilizadas na pesquisa possuem a mesma limitação, pois todas elas utilizam uma única formatação gráfica.

Por outro lado, a análise da pesquisa revela o anseio do usuário de ter mai informações por meio de símbolos/ocorrência de códigos, tendo inclusive sugerido os seguintes: a) explicar como interpretar, b) novos tipos de visualização gráficas, c) mostrar todas as características, d) alterar codificação visual, e) novos tipos de visualização de dados.

Então, de acordo com os resultados das entrevistas, acredita-se que simples intervenções visuais possam acarretar em uma melhor compreensão dos resultados das técnicas de inteligência artificial explicável.

7.2 Recomendação de melhoria das técnicas de explicabilidade

De acordo com os resultados obtidos na pesquisa com a utilização da Teoria fundamentada em dados, as técnicas de explicabilidade podem ser melhoradas com as recomendações abaixo listadas:

7.2.1 Melhorar a visualização da informação em sistemas de IA explicável

A informação quando visualizada de forma clara e correta permite a exploração da capacidade humana de processamento visual e tal recurso é um agente facilitador para o entendimento dos resultados nas técnicas de IA explicável.

Portanto, verifica-se que mostrar a técnica de explicação com as características de maior/menor influência pode aumentar a explicabilidade.

Do mesmo modo, demonstrar na técnica de explicação todas as características da predição resulta em maior compreensão e em maior grau de visualização da técnica de explicação para a explicabilidade.

Em que pese as técnicas de IA explicável sejam consideradas por muitos como não explicativas, tal percepção pode ser modificada à medida que melhorias visuais sejam introduzidas a essas técnicas, como por exemplo:

• Implementar outros tipos de gráficos;

• Incluir informações de ajuda (help) ou informações de "como Interpretar os dados";

• Incluir visualizações de dados avançadas como, por exemplo, mapas mentais;

• Permitir a alteração da codificação visual, tais como: o tamanho dos elementos, cores, saturação das cores, formato e textura;

• Permitir a visualização de todas as características do modelo subjacente.

7.2.2 Incluir mecanismos de rastreabilidade nas técnicas de IA explicável

Importa dizer que a rastreabilidade da informação permite compreender os relacionamentos existentes entre os dados ou entre artefatos, arquitetura e implementação.

Portanto, considerando que ‘xAI não é explicativa’, então mostrar na técnica de explicação todas as características da predição pode melhorar a compreensão do resultado e, nesta esteira, a inclusão de mecanismos de rastreabilidade na técnica de explicação

aumenta ainda mais a sua explicabilidade.

Desta forma, sugere-se a inclusão de mecanismos de rastreabilidade nas técnicas de IA explicável, tais como:

• Incluir peso de contribuição das características na explicação em medida conhecida;

• Agrupar características por peso (hierarquia);

• Exibir todas as características do modelo subjacente;

• Exibir os mapeamentos da entrada, processamento e saída das características mais importantes;

• Emitir símbolos ou regras juntamente com sua saída específica a fim de auxiliar no processo de compreensão da lógica por trás dos mapeamentos feitos.

Uma vez implementadas as melhorias apresentadas acima, acredita-se que a rastreabilidade nas técnicas de IA explicável se aproximará dos conhecidos sistemas compreensíveis que, por definição, são: “aqueles que emitem símbolos, permitindo explicações orientadas pelo usuário sobre como chegar a uma conclusão, ou seja, compreender por que uma certa saída está associada a uma certa entrada” [61].

8. Conclusão

Inicialmente, é importante mencionar que este estudo tem por fundamento conceitos da literatura sobre: inteligência artificial, aprendizagem de máquina, inteligência artificial explicável e suas respectivas técnicas.

Além disso, registra-se que o estudo aqui apresentado avaliou a explicabilidade de técnicas de inteligência artificial explicável na perspectiva dos especialistas em domínio (médico oncológico). Embora existam outros estudos comparando técnicas de inteligência artificial explicável [66–68], esses não focaram nas necessidades cognitivas específicas dos diferentes consumidores de explicações, mas focaram nos problemas de visualização, desempenho e auditoria do modelo.

Consigna-se, ainda, que com base em entrevistas com oncologistas (especialistas em domínio) interagindo com diferentes técnicas de IA explicável (SHAP, LIME e Permutation Importance), identificamos alguns aspectos importantes que ainda precisam ser abordados. Concluímos que as técnicas de IA explicável são informativas e não explicativas, limitadas à exibição das características mais importantes na predição, em vez de permitir a exploração de explicações em diferentes perspectivas.

Nossa pesquisa se concentrou na perspectiva do especialista em domínio no contexto de compreensão e aceitação de uma explicação baseada em IA explicável, enquanto as pesquisas atuais se concentram em IA explicável numa perspectiva dos projetistas e especialistas em inteligência artificial.

E, por último, avaliou-se a percepção do usuário especialista no domínio médico quanto a sua confiança na inteligência artificial como instrumento para obtenção de resultados que possam influenciar na sua tomada de decisão especializada.

8.1 Contribuições do Trabalho

A principal conclusão deste trabalho como contribuição para o fortalecimento e, por consequência, implantação do sistema de inteligência artificial explicável como ferramenta para encontrar melhores e mais seguros resultados advém da constatação de que o usuário especialista no domínio médico não consideram as técnicas de inteligência artificial explicável explicativas e, por essa razão, são pouco úteis na clínica médica.

Portanto, a contribuição maior deste trabalho visa demonstrar, de forma científica, as técnicas utilizadas na pesquisa e como elas fundamentam os seus resultados, de modo a ampliar o conhecimento da comunidade.

Com a disseminação do conhecimento a partir deste estudo, a comunidade de inteligência artificial tem condições de elucidar suas demandas relacionadas à explicabilidade e à compreensão dos sistemas de inteligência artificial explicável (XAI) e, por conseguinte, confiar na inteligência artificial.

E, mais, constatou-se que o presente estudo permitiu que 03 (três) populares técnicas de inteligência artificial explicável fossem direcionadas a especialistas no domínio médico oncológico no contexto brasileiro, o que é inovador.

Por fim, entendemos que este trabalho é um primeiro passo importante para realmente criar diretrizes e um conjunto de boas práticas de projeto de sistemas de inteligência artificial explicável com o fito de atender as necessidades explicativas dos usuários especialistas no domínio.

8.2 Restrições

Como importantes restrições desse trabalho podemos citar o não acesso a um conjunto de dados mais robusto, com um alto volume de dados, além de não serem um conjunto de dados com informações da população brasileira.

Nos limitamos a usuários especialistas no domínio médico e poderíamos ter aplicado a outros domínios críticos como mercado financeiro, por exemplo. Além de comparar com especialistas de outros domínios considerados não críticos e outros tipos de usuários, como: agencias reguladoras, pessoas afetadas pelo sistema, gestores e pacientes.

8.3 Trabalhos futuros

A partir do presente estudo, propõe-se para o futuro avaliar as técnicas de inteligência artificial explicável sob a perspectiva não só da comunidade médica especializada, mas, também de outros domínios e usuários, tais como: leigos, especialistas no domínio/ especialistas em T.I., bem como em diversos domínios (domínio não crítico, domínio intermediário e domínio crítico).

A propositura relativa à ampliação do círculo de destinatários da pesquisa tem por objetivo reforçar os seus resultados e, assim, desenvolver novas teorias que permitam ampliar o conhecimento da comunidade de inteligência artificial explicável.

Além disso, propõe-se a implementação das recomendações de design para os sistemas de inteligência artificial explicável agnósticos para melhor comparação e prova dos resultados da pesquisa.

Por fim, a criação de metodologia qualitativa para avaliar o grau de interpretação humana das explicações encontradas para os modelos de aprendizado de máquina.

A. Entrevista PO01

Data: 27/01/2020

Onde: INCA

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Os que mais influenciaram né? Todos aqui tem um peso, mas os que mais influenciaram, eu acredito que seja a idade, a citologia, o número de parceiros sexuais, número de gravides e a idade da 1º relação sexual e o teste de Schiller.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?"

Com esses fatores que mais pesaram, sim. Fazem sentido. Para esse caso em especial. Esses primeiros aqui que eu elenquei com os que fazem mais sentido, são os que traduzem realmente nos de risco aumentado para câncer de colo de útero. Em um primeiro olhar faz sentido, não sei se eu confiaria, porque é necessário mais dados, outras análises, enfim. Ele pode ter acertado nesse, mas eu não sei quantas vezes ele acerta, eu não vi outras análises, se coincidem o diagnóstico com outras análises. Nesse momento ele acertou aqui, mas não sei reprodutibilidade disso.

Concordo parcialmente, porque não necessariamente o paciente que tem essas características, vai ter um risco alto de câncer de colo de útero ou vai ter um diagnóstico de câncer de colo de útero. Tem um risco alto dela ter, mas não traduziu em diagnostico para

ela. Depende da biopsia mesmo, que o padrão ouro é a biopsia.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada nestes métodos?

Que os valores próximos serem agrupados. Ao invés de fatores separados, a partir daqui, fossem feito grupos de fatores, como se colocassem em ordem mesmo. Essas características no grupo 1, essas características da paciente no grupo 2, características no grupo 3. Até para criar fatores de maior peso, fatores intermediários e fatores de menor peso, isso ajudaria bastante o diagnóstico.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Que essas características aqui, o Teste de Schiller, Colposcopia, a citologia, número de parceiros sexuais, idade. Essas características são as que aumentam a probabilidade dessa paciente ter câncer de colo de útero. Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?"

Permite, porque te mostra. Na verdade, dá para você entender o diagnostico afirmado de câncer de colo de útero, justamente pelos fatores de risco que ele apresenta. Os maiores fatores de risco aumentam a probabilidade de ela ter câncer de colo de útero. Nesse momento não dá para confiar, porque eu não tendo outras análises que ele tenha feito. Com um caso só, se somam como coincidentes né. Precisaria de vários casos, caindo na mesma análise, para ver o percentual de acero, para poder confiar. Eu utilizaria como auxílio preditor.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Da mesma forma que o outro método, agrupar as caraterísticas pela probabilidade. Para você criar como se fosse uma hierarquia de grupos de características. Se ela apresenta essas características que tenha um peso maior, que essa que um peso intermediário, que essa que tem um peso menor. E quanto isso se relaciona com outras características.

Por exemplo, se você juntar duas com um peso menor, daria uma intermediária? Ou duas com um peso menor, daria uma com um peso maior? Ou uma de peso intermediário com uma de peso menor, daria uma de peso maior? Enfim, dar uma relação desses pesos.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Aqui é como se ele ordenasse do fator que mais pesa para o fator que menos pesa para ser um preditor, de ela ter uma probabilidade de ter câncer de colo de útero. Ai conforme você vai andando na réguinha, isso vai pesando mais ou menos. É visual né. Não é numérico como o outro é.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O quanto você concorda com a seguinte afirmação: Ä explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?".

Sim. Justamente porque ele me diz as características que aumentam o risco de ela ter câncer de colo de útero. Ele me mostra essas características, baseado em todos os dados, onde os dados são mais frequentes em quem tem câncer de colo de útero. Essas características. Eu confio com cuidado, o diagnóstico você tem que ter um exame de certeza. Então eu confio com uma grande desconfiança. Mas não me dá o diagnóstico. Entendeu? Concordo parcialmente, justamente pela falta do teste padrão com comprovação do diagnóstico. Na verdade, ele me da uma probabilidade, uma alta suspensão, mas não me diz. Como nem todo suspeito é culpado, se realmente tem o diagnostico de verdade.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

A comparação com outros pacientes do mesmo perfil, que tem o diagnóstico de câncer de colo de útero. Esse aqui faz, na verdade ele ordena pela importância, pelo peso. Nessa reguinha ele ordena pelo peso, é bem visual. Tanto que eu não tive nenhuma dificuldade de visualizar a ordem de importância, coisa que nos outros dois métodos eu tive dificuldade. Nos outros dois métodos eu tive que me reportar aos números, esse aqui foi só olhar a reguinha.

Qual método de explicação você mais gostou? Por quê?

Esse, o SHAP. Ele é muito mais visual.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Como apoio sim, não como diagnóstico definitivo.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Depois de estabelecido essa capacidade de explicação, submeter esses modelos de explicação a trabalhos de concordância, confiabilidade, análise de resultados, análise de métodos de diagnostico.

.

A. Entrevista PO02

Data: 27/01/2020

Onde: FIOCRUZ / IOC

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Eu explicaria que o que mais influenciou na predição foram características relacionadas ao exame de Papanicolau, teste de Schiller, assim como utilizar contraceptivos. Na verdade hormonal e em Hinselman, só que eles tiveram o valor baixo e que o valor preditivo foi um valor muito bom.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Exame de Papanicolau, Teste de Schiller e Número de parceiros sexuais.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?".

Sim. Os descritores que mais influenciaram no valor preditivo.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada nestes métodos?

Nossa é difícil. Acho que ele está bem explicativo. O peso de repente, porque aqui ele está mostrando os que pesaram mais, mas não o peso.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo

sistema "Dr.Inteligência Artificial"pelo método SHAP?

Então, o outro eu acho mais intuitivo (LIME). Porque ele mostra bem os que tiveram um peso maior, na verdade os que influenciaram mais... mas o que tem embaixo a gente não consegue ver. O outro eu consigo ver os descritores que estariam relacionados a uma pouca influência Esse a gente não consegue, esta exprimido aí, mas eu acho que ambos. Agora nesse ponto aqui é o número absoluto? Então por isso que o peso é importante, eu teria o limite, por exemplo um parceiro sexual. diagnóstico

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Os mesmos.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?".

Sim. Então mesmo para esse, quanto para o outro, eu acho que a resposta seria a mesma. E ai eu acho que tem a ver com o background da pessoa, ou seja, tem a ver com expertise de quem está avaliando. Então, para um especialista, um oncologista ou ginecologista, eu acho mais fácil de entender do que um público leigo. Porque um paciente olhando isso aqui, o número de parceiros, o exame, para ele não fica muito claro. Mas quem entende um pouco da doença, consegue identificar bem, então para o especialista está excelente.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Os pesos também.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Estão muito semelhantes né. Embora aqui na verdade os dados são muito parecidos. Se eu tivesse que escolher um dos três modelos, por exemplo, não sei se é uma das perguntas né. Mas eu acho que esse está bem descritivo o resultado. Então estar mostrando o que pesou mais e o que pesou menos, todos mostraram. Mas eu acho que esse está com um número, não sei se é a forma de mostrar dele, mas o resultado está mais claro. Esse

está com uma forma de representação mais clara.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

As mesmas.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?"

Sim. Como especialista sim, novamente, para especialistas sim. Acho que são todos muito parecidos. Acho que o conhecimento das características. Nesse caso específico você usou os mesmos descritores né, as mesmas características né? E os três conseguiram recuperar o que tem um peso de importância na doença, então eu acho que isso é o que mais se destaca. Mas os três foram muito semelhantes.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Acho que todos eles o peso. Que é o limite né. Porque pesou mais o Teste de Schiller? Porque deve ter uma propensão, um limiar de tanto a tanto, que tem uma característica que leva ao câncer. Então marcadores que são importantes, valores que são importantes.

Qual método de explicação você mais gostou? Por quê?

Esse, o Permutation Importance. Eu gostei bem dos três. Mas eu achei o Permutation Importance mais claro, porque tem esse tamanho aqui, com os ranges de desvio, e ai quando você não tem um peso, você entende onde estaria. Se fosse por exemplo uma clusterização, você conseguiria onde aquele grupo se encaixaria. E visualmente é mais fácil. Eu não sei se é para mim que já trabalha com alguma coisa de inteligência artificial, mas para uma pessoa leiga o 3º talvez seria mais fácil.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Super confio e uso bastante inclusive.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Sim, com certeza. Tanto para descritores utilizando dado de imagem, quanto para

RNA, DNA e dados de imagem.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

A. Entrevista PO03

Data: 27/01/2020

Onde: FIOCRUZ / IOC

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Pelo gráfico, ele tem uma apresentação de probabilidade né, onde a gente consegue visualizar e colocar em percentual quais são as chances de a pessoa desenvolver um câncer. Tem até em função de ter as cores, com as características e os valores, eu acho que isso acaba facilitando um pouco. Eu acho que destaca bastante os testes utilizados, os exames que são utilizados hoje, que são coisas muito corriqueiras e muito comuns em mulheres, fáceis de responder. E eu acho que isso auxilia para a gente ter depois um dado mais, eu não digo robusto, mas um dado que o conjunto dele dá uma robustez no resultado.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

A idade que tem o valor aqui mais alto, a idade da primeira relação sexual, os hormônios aqui nem tanto, o Teste de Schiller também aqui.

Os hormônios são engraçados que eu achei que poderiam influenciar mais, mas eles influenciam pouco né. O Número de parceiros sexuais, também influenciam bem, quer dizer, um número de parceiros. Bem interessante isso.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?".

A gente tem que levar em consideração que você está com uma pessoa que tem um

certo entendimento e um grau de instrução. Então a gente tem que ir para os dois lados. Como uma pessoa de nível um pouco mais primário de intelectualidade, de aprendizado, veria isso? A meu ver está claro, mas isso porque eu conheço um pouco desses dados de análise, então fica mais fácil para eu entender. Mas vamos supor minha mãe que não tem conhecimento de nada disso, ela não entenderia. Entendeu? Para mim não tem dificuldade, eu entendo um pouco, trabalho na área, tenho amigos que trabalham na área, fica mais fácil de entender. Mas para um leigo, uma pessoa que nunca viu, eu não sei se seria fácil. O resultado talvez seja fácil, mas a compreensão do que está sendo apresentado aqui, para um leigo, para a população em geral, não sei se fica tão fácil assim de entender.

A sua apresentação previa, o meu conhecimento prévio, acho que tudo isso ajuda. O fato de eu já ter um grau de estudo um pouco maior, de entender, de já ler, de entender um pouco mais de informática, de TI, de conhecer um pouco melhor esses programas. Isso ajuda muito. Agora se você tivesse que fazer essa validação numa dada população em geral, eu não sei como seria. Mas para mim foi um pouco mais tranquilo.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada nestes métodos?

Acho que um gráfico ajuda. Quando você coloca as características, com os valores, e um gráfico de barras onde tem os valores no gráfico, acho que isso faz toda diferença. Até para um leigo fica mais fácil de visualizar. Uma apresentação com um gráfico ou com um fluxograma, dependendo do que você quer explicar ou como você vai colocar a sua metodologia. Quais são os métodos que v que você utilizou para fazer a análise, eu acho que isso ajuda.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Esse modelo foi um modelo que eu gostei muito quando você explicou para a gente, porque ele ranqueia quais são os testes, que tem maior peso, dentro do dado como um todo. Para você chegar a sua conclusão se a pessoa tem maior ou menor predisposição, ele faz um ranqueamento de parâmetros. E esse ranqueamento, ele corrobora em parte com esses dados, um pouco do nosso estudo em câncer. Por exemplo, o fumo está diretamente correlacionado com o câncer, o aumento da idade também está diretamente relacionado com o câncer né. Mas tem outras coisas que está especificamente ligada ao câncer de útero, que é o Papanicolau, os contraceptivos em anos, que descreve melhor e ele ranqueia em peso melhor do que o LIME. Então eu acho que ele da um peso maior para o seu dado,

entendeu? Eu gostei muito dele, o ranqueamento que ele colocou, os parâmetros que ele utilizou. Eu gostei bastante.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O Teste de Schiller, a Idade. Para qualquer tipo de câncer a idade está diretamente relacionada, quanto maior é a idade, maior a chance de desenvolver câncer. E o que me chamou muito atenção, que bate muito com o LIME, é o número de parceiros sexuais. Muito interessante. Tem também aqui uma coisa que me chamou muito atenção, que é a condilomatose perineal, que é uma doença comum, principalmente em pessoas de baixa renda, e as vezes a pessoa fica anos sem saber que tem. E também está relacionada com tumor de câncer de colo de útero, e não tem tanto peso quanto o teste de schiller e a idade, mas também está relacionada. Bem interessante. As doenças inflamatórias também, como qualquer processo inflamatório, independente do órgão, você tem que tomar cuidado. Então aqui ele também destaca a inflamação na região pélvica, que também pode ajudar no desenvolvimento. Se for uma coisa crônica, uma inflamação crônica, na região uterina por exemplo, pode levar a um desenvolvimento futuro de câncer. Então o processo inflamatório nessa região, então ele destaca pontos, que são importantes para o cuidado, que de repente num processo de publicidade para divulgar o cuidado com a mulher, como a mulher se previne de doenças como essa, a gente pode prevenir com cuidados com esse. Cuidado com as doenças, doenças inflamatórias, que são coisas que a mulher pode se cuidar e as vezes deixam pra lá. Fazer os testes como Papanicolau, que é fundamental para prevenir o câncer de colo de útero e por lei deveria ser feito uma vez por ano

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente".

Permite.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada nesse método?

Ai eu volto a dizer né, o peso, essa estrutura poderia estar em gráfico. Eu acho que poderia colocar, por aqui é um ranking, ou as características principais numa forma de 1º, 2º e 3º com o peso dele, para ficar mais claro para uma população leiga. Ou em gráfico. Ou num fluxograma, que também ajuda. Toda forma que você consegue melhorar o visual

para a pessoa entender melhor o que esta escrito, ao invés de uma tabela. Tabela nunca é didática, não fica claro. Entendeu? A gente mesmo fez um curso, que tudo que a gente divulga, tem que ser de fácil entendimento para a população. Então quanto mais claro isso, melhor. A tabela nunca é clara apara a população que é leiga. Entendeu?

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Esse eu gostei. Esse é o que ficou com o visual mais claro para uma pessoa entender. Esse você consegue entender que essa paciente tem 81% de chance de desenvolver câncer de colo de útero. E aí ele detalha aqui embaixo, quais são as características principais que levaram esse programa a chegar a esse percentual. Então ele mostra, isso é o que? Isso é o visual. Faz toda a diferença, não é uma tabela. Ai se você pega isso aqui e melhora, ou seja, tira as informações ne, data information, de TI e coloca isso numa apresentação com esse desenho, porque aqui já está em português. Aí você coloca o que é maior predição, baixa predição, de uma forma clara, para uma pessoa leiga entender, ela fica perfeita. Fica perfeito para entender. E aqui ele mostra, quais foram os descritores que determinaram esse valor. Então o Teste de Schiller está em 1º lugar, o exame de Papanicolau vem em 2º, o número de parceiros, o que corrobora com os outros. Dos programas que você utilizou para fazer esse tipo de análise, todos eles colocam o número de parceiros sexuais né. E a idade da 1º relação, quanto mais cedo, mais chance de desenvolver um câncer. O visual aqui ele faz toda a diferença, fica de mais fácil entendimento.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, Papanicolau e Número de parceiros, Idade da 1º relação sexual.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente?".

Sim. Ajudou bastante. Essa questão da imagem né, doa ranqueamento, ele destaca quais são os foram utilizados como parâmetro, que influenciaram mais para chegar a esse percentual. Eu achei ótimo.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Eu acho que a única coisa que falta aqui, é dizer que isso não significa que a pessoa vai ter câncer, mas que é um percentual que prediz a chance de desenvolver. Mas se a pessoa tiver os cuidados de fazer os exames, de se cuidar, previamente fazer os exames preventivos. As pessoas vão no mínimo descobrir na forma precoce da doença, tratar de forma rápida e não desenvolver a doença de uma forma agressiva. Então eu acho que isso ajuda muito. Então de repente deixar aqui bem claro que é uma predisposição e não um fato né, e que precisa fazer exames periódicos e se cuidar. Isso vai ajudar muito aquelas mulheres que fazem exame de forma preventiva, então eu acho que isso ajudaria bastante.

Qual método de explicação você mais gostou? Por quê?

O SHAP. Então, justamente da forma como ele apresentou os parâmetros, ele já mostra bem claramente o percentual de predisposição ao desenvolvimento do câncer e ele já indica quais são os principais parâmetros que levaram ele a chegar a essa predisposição de percentual. Então eu o achei melhor, achei muito claro, muito objetivo.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Sim, eu acredito. Acho que é uma forte ferramenta, entendeu? Aqui ele mostra muito bem isso, é uma ferramenta que a gente pode utilizar principalmente nos dias de hoje, com a vida corrida que a gente tem. A IA vem trazer de uma forma mais simples, soluções ou questões, que fazem a gente ou abrir os olhos e ficar mais atentos a questões como essa, o desenvolvimento de um câncer.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Eu acredito, mas acho que existem muitos passos pela frente a serem alcançados nesse caminho. Porque existe dois grupos distintos. Aquele que ainda são daquela mente antiga né e que não acreditam nas ferramentas tecnológicas e que acreditam somente no contato direto com o paciente e não tem ainda muita fé nesse sistema. Mas tem hoje os jovens pesquisadores, os médicos mais jovens, que levam isso muito a sério e eu acho que são eles que vão colocar toda essa tecnologia nova para frente. E com esses que temos que agarrar e mostrar que é uma tecnologia boa, que é uma ferramenta de inovação, que pode contribuir muito para o diagnóstico precoce do câncer. Ou não só para o diagnóstico, mas também para o acompanhamento para aqueles que não tem câncer, mas que tem uma forte predisposição. Então são com esses profissionais que temos que contar, porque os antigos que não acreditam nisso como ferramenta não vão mudar. Mas esses que estão começando agora e identificam isso como uma importante ferramenta, a gente tem que mostrar e andar junto com eles.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

A gente tem que mudar a forma. Por exemplo, a forma como você me apresentou hoje, o médico não vai aceitar. Por quê? Porque não está claro, não está obvio. É o que eu estou falando, nem todos conhecem a linguagem da tecnologia da informação. Então tem que colocar de uma forma que para eles fiquem claro, que é uma validação, para uma avaliação de predisposição ao câncer a possíveis pacientes. E aí mudando essa formatação, que não está muito clara. Tem que melhorar essa forma, melhorar a apresentação. Melhorar a página, melhorar o layout. Para de fato ficar numa linguagem mais fácil e de fácil compreensão para eles. Porque quem vai acabar aceitando isso é o médico, então tem que ser colocado de uma maneira fácil, de uma linguagem mais fácil, para que ele consiga entender e aceitar, e fazer as perguntas para a paciente dele, porque é ele que vai aceitar.

A. Entrevista PO04

Data: 27/01/2020

Onde: FIOCRUZ / IOC

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

É, o que me parece é que os dois testes e exames de rastreamento né, tanto o Schiller quanto o Papanicolau são bastante importantes para predizer a predisposição de ter ou não esse tipo de câncer. O que eu acho muito válido e relevante, ele ter dado esses 2 como os principais, depois ele vem com o número de parceiros sexuais e com a idade da primeira relação sexual? Isso? Também acho bastante relevante visto que é câncer de colo de útero, a idade da primeira relação sexual eu não sei, não consegui ver nenhuma relação tão direta. Não entendi tanto o quanto esse dado está ajudando, mas os outros bastante.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, exame de Papanicolau e Número de parceiros sexuais

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Se ajudou a entender o diagnóstico? Sim, mas não sei se ajuda a compreender. Tá, acho que eu entendi o que você quis dizer. Se você tem uma um dado baseado em Inteligência artificial que diz, você tem 81% de chance de ter câncer, esse dado me explica sim do porquê que você chegou a esse percentual. E também gostei desse sistema de cores. Tanto de cores, quanto de tamanho das barras para explicar o quanto que são mais

determinantes. As cores e tamanho das barras.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Talvez aqui se o Teste de Schiller foi positivo ou não, ele está rosa porque ela fez o teste. Mas eu acho que dizer, olha você tinha dado positivo ou negativo e de fato ela tem câncer ou o contrário né, o teste deu negativo e ela não tem. Acho que seria legal. E outra coisa na pergunta anterior, eu acho que ter no texto o que é cada barra, é muito importante também. Porque você poderia ter só uma sigla, alguma coisa, eu acho que assim fica mais claro, o texto inteiro do que significa cada barra.

Além do seu positivo e negativo? Bom, eu acho que além do que você já mostra, exatamente isso, se é positivo ou não e aqui você tem uma informação que eu estou imaginando o tamanho da barra, diz que é de 0,4 a 30 e poucos por cento tá prevendo a partir do teste de Schiller. Talvez botar aqui o percentual que cada um contribuiu para essa predisposição no final. Acho que ficaria legal desse 81%, tantos por cento são advindos do lado de teste de Schiller, tantos por tanto do outro e 10% são desse. Aí você tem como dizer o quanto que o número de parceiros sexuais por exemplo, ou de cada um deles, tá contribuindo para o resultado final. Acho que seria interessante, ter de alguma forma, talvez aqui embaixo, um valor em percentual.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Esse já é um pouco mais complicado de ser interpretado que o anterior (SHAP). O que eu consigo imaginar é que ele te dá de fato uma ordem de relevância ou de força que cada característica tem para o resultado final. Mas esse peso né, em 0,0195 me diz menos do que o gráfico anterior, do quanto que ele foi relevante para chegar ao resultado final. Então eu entendo que é uma ordem de relevância de cada um deles tem para o resultado final, mas eu acho mais confuso. Apesar de outras pessoas poder achar mais fácil a tabela do que outro, mas eu ainda prefiro o outro. A não ser que você consiga transpor do peso para o percentual de que foi importante para o resultado final. Aí talvez ficaria mais fácil.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O teste de Schiller, a idade e o Exame de Papanicolau.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Com certeza. Eu acho que os descritores né. Você ver um número grande de perguntas que você faz para o sistema. Não sei se assim né, se exatamente você pergunta. Então se a paciente apresenta ou não aquela característica. Se apresentasse sei lá 5 características, eu acharia mais fraco. Mas como você tem um número grande de descritores, eu acho que fica bastante confiável o dado final. E você tem uns que são zerados né, ou seja, eles contribuíram pouco. Esse 0 significa que contribuiu pouco ou que ela não tem? Então eu acho que é isso mesmo.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada nesse método?

Eu acho que isso que eu tinha falado inicialmente ne. Ter um percentual ou uma força que cada um desses teve no resultado final, que eu acho que é esse peso, mas eu acho ele pouco explicativo. Eu não sei o peso final seria 1 e a soma de todos eles seriam 1.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Então vamos lá. Vamos por parte né. O método foi capaz de acertar em 81% das vezes e errou em 19% das vezes, o que me parece bastante forte. E aí você esmiuçando isso, que os preditores principais foram o Teste de Schiller que ele foi 100% capaz de prever, o exame de Papanicolau também foi 100%. É isso? E aqui um pouco do cottof que eu imagino. A idade da primeira relação sexual foi 15 anos e a idade dela atual era 21 anos. Era isso? Eu acho que dá para entender bem.

Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, Exame de Papanicolau e Idade da 1º relação sexual

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Então uma pergunta, o sistema e ele dá um gráfico dessa forma ou ele gera uma tabela

dessa forma? Então os que estão em azul foram os que menos influenciaram né? Talvez o que faltasse aqui era você ter uma legenda das cores, do laranja e do azul, para mais influência ou menos influência e também de repente até uma gradação dentro do laranja e do azul. Mais sim, dá para entender sim. Depois de ter essa sua explicação de que o azul é o menos influente e o laranja mais influente. Ele é menos influente ou está mais relacionado a não ter câncer? Então ele influencia a não ter câncer. Eu acho que primeiro esse tipo de gráfico que é apresentado aqui, de ter a barrinha do lado direito com câncer com a cor bem clara e não câncer para o lado de cá, foi legal. E aí fica claro quais são os descritores que estão relacionados a não ter câncer nesse caso e quais que foram relacionados a ter câncer também nesse caso e também essa tabela dessa forma dividido em cores. Acho que foi legal

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Esse aqui é o resultado de um caso né? Aqui por exemplo que eu não entendi né, que ela tem 0 gestações, então não ter tido gestação, tá mais implicado a ela desenvolver câncer. Isso eu não entendi. Esse eu não sei o que você poderia implementar. Esse eu não sei. O que eu gostei desse aqui é que ele vai numa crescente, ele mostra o total, e mostra o que cada um contribuiu. Achei ele bastante interessante por isso. Acho que isso foi mais completo em relação aos outros. Achei esse melhor que o primeiro.

Qual método de explicação você mais gostou? Por quê?

Do primeiro (SHAP) e o terceiro (LIME). Do SHAP e depois LIME. O primeiro porque eu achei esse tamanho da barra, eu acho que ele te dar a questão das cores que já diz o que é mais relevante para predisposição ao câncer e te dar o valor em percentual de predisposição, o tamanho da barra e ainda da a legenda de cada um, então eu achei bastante claro na mensagem final. Apesar de não ter os baixos né, aparece pouco, mas eu achei ele muito interessante por isso. Eu achei ele bem claro.

O segundo (Permutation Importance) que é aquela tabela, talvez se tivesse mais uma coluna naquela tabela já resolveria. Mais uma coluna com o percentual ao invés daquele peso, talvez ele fosse melhor. Porque ele é bem claro, aquela tabelinha é simples e te dar uma informação super relevante. É claro que ali são os primeiros que são os mais importantes e os últimos são os que não contribuíram para o resultado final. Então talvez se ele tivesse só essa tabela, seria um forte.

Esse aqui é legal (LIME), porque ele vai em três camadas de explicação, mas eu achei

ele o mais complicado de entender.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Com certeza. Tenho bastante.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Eu acredito que a revolução vai vir por aí. Nós temos muita informação, muitos bancos de dados e essa informação ela é pouco explorada. Então eu realmente acho que a revolução, no diagnostico principalmente, na decisão de que tipo de medicamento ou outro, vão vir da IA. Com certeza. E por isso eu te parabenizo por estar estudando isso, ainda mais o câncer. Você está de parabéns.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Bom, eu acho que para você convencer os médicos de utilizar a inteligência artificial, eu acho que primeiro tem que ir fazer a validação de estudos para outros tipos de câncer, outros tipos de bancos de dados, mais complexos inclusive. E aí você sempre tendo esse valor alto de predisposição, você convence. Então eu acho que é o normal, você tá indo por um caminho, tá fazendo com um tipo, um banco de dados. Acho que teria que validar em outra corte, em outro tipo de amostra né, de paciente. Para esse tipo tumoral e depois expandir para outros tipos tumorais.

Porque eu acho os médicos muito céticos, principalmente pela sua informação inicial, é uma caixa preta e eles não tem como saber se é confiável ou não. Então eles entenderem o porquê que o sistema prediz.

A. Entrevista PO05

Data: 30/01/2020

Onde: Hospital da PUC-RIO

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Aqui o Teste de Schiller e Papanicolau da mais chance de risco do que o número de gestações e a Colposcopia por exemplo. E assim, na verdade a paciente ter feito Papanicolau, só me diz uma coisa quando você diz ter feito Papanicolau é ter feito o exame ou ter dado positivo? O exame de Papanicolau com alteração? O exame de Papanicolau é o 1º exame de rastreio, fora a captura hibrida da detecção HPV, então o Papanicolau dando positivo, realmente é um fator de risco grande, agora a Colposcopia sendo negativa, é um exame mais específico que esses dois. Entendeu? Então aqui se a Colposcopia for negativa, eu levaria mais em consideração esse resultado negativo aqui para câncer, que a Citologia isoladamente. Ai eu não sei, como vai ser isso. Porque na verdade são técnicas complementares né. O exame de entrada é preventivo, quando alterado, encaminha para colposcopia e na Colposcopia a gente faz o Teste de Schiller. Ai se tiver achados, aí a gente conduz de acordo com os achados, se a Colposcopia vier normal, nós já afastamos a chance de câncer de colo de útero ou doenças pré-câncer. Aí depende do achado na Colposcopia. Sim. A técnica ajudou sim.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Para mim mais o Papanicolau e o Teste de Schiller. Porque a número de parceiro sexual, idade e idade da 1º relação sexual a gente não leva tanto em consideração. Porque a infecção quando ela é na juventude, ela tem um caráter mais transitório do que

quando ocorre numa idade mais avançada. A idade a gente não leva tanto em consideração, número de parceiros também não, porque é uma infecção de grande incidência na população mundial. Então quase todo mundo tem HPV, e pra gente não interessa. O que interessa mesmo é o segmento com o preventivo e os demais exames de segmento mesmo, entendeu?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

De previsão? Sim. Sim. Da uma ajuda muito grande. Agora você usaria esse modelo em lugares de mais difícil acesso ou na população em geral por exemplo? O que mais me fizeram entender? A visualização rápida do que realmente é relevante em termos de fatores predisponentes para câncer e doenças pré-câncer. Num único olhar é possível ver graficamente o que é mais relevante ou não.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Eu acho que facilitaria dizer qual o tipo de alteração tem nesse exame de Papanicolau. Porque só tem câncer de colo de útero em quem apresenta doenças pre câncer do colo. Por exemplo uma lesão de baixo grau ou uma lesão mais especifica ela não fala tão a favor para uma doença do colo de útero, eu não levaria tanto em consideração, como se fosse uma lesão de alto grau ou outras alterações. Como eu te falei, a presença de captura hibrida, Teste de Schiller, dependendo da associação com a Colposcopia eu levaria mais em consideração o exame de Papanicolau. O que que deu nesse exame? Qual foi o resultado dele? Não sei o que deu aí, qual foi alteração? Tem uma gama de coisas que podem acontecer e nem tudo significa doenças pré câncer. Isso deixaria um pouco mais confiável.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Eu explicaria que tem mais predição de câncer de colo de útero o que estão em laranja. É isso que eu vejo.

Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, Colposcopia e Papanicolau.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Câncer não, de doença pre câncer né? Permite sim. São plausíveis, são possibilidades. Aí entra outra questão, como eu te falei, quando são DSTs, aí entra HIV? Ahh aqui ela não tem. Não é isso que você botou? A gente sabe que tem DSTs que influenciam a infecção por HPV. Mas eu acho que poderia desmembrar essas DSTs para as que gerem imunossupressão, que é o HIV né. Acho que seria interessante colocar isso a parte, separado ne. Mas está bom. Graficamente e visualmente é possível identificar os fatores de maior probabilidade para ajudar no risco de câncer.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Como eu te falei, colocar separadamente a questão de infecção por HIV, que é um dos fatores para imunossupressão. Colocaria também doenças imunossupressoras que tem bastante associação. E colocaria qual o tipo de alteração no Papanicolau que surgiu. Por exemplo, doenças de baixo grau e doenças de alto grau. Dar por exemplo a periodicidade do último exame preventivo. Se ela fez o exame há três anos ou há cinco anos, fica mais confiável de saber se ela tem predisposição numa paciente que você tem um acompanhamento mais periódico, do que uma paciente que você não acompanha há dez anos por exemplo né. Também aumentaria o risco de câncer para uma pessoa que não está sendo acompanhada né.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Deixa-me dar uma olhada aqui. Esse não faria tanto sentido. Esse pra mim o resultado não fez tanto sentido não. Já não traria para mim tanta alusão não. Ele põe por exemplo o Teste de Schiller mais importante do que o exame de Papanicolau, então para mim é questionável.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, Idade da 1º relação sexual e número de gestações. Foi o que eles

colocaram como principais.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Não. Para mim não foi tão alusivo. Não foi tão explicativo. Eu acho na verdade de todos, o que eu mais gostei foi o segundo (LIME). Eu acho que didaticamente não é uma coisa que olha e seja auto explicativo, se é pela ordem de colocação, se é pelas numerações e não faz tanto sentido para mim alguns dados serem antes ou depois. Diferente dos outros gráficos assim, eu não me senti confortável de olhar e explicar o resultado com essa tabelinha ai não.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Eu acho que eu senti falta de explicar como deve ser interpretado, se é pela ordem de aparecimento, se é pelos valores. Ai eu não sei como avaliar, se é pela ordem, se é pelos valores dispostos. Não ficou muito claro. Acho que colocaria uma legenda, uma explicação clara de como deve ser interpretado. De repente isso.

Qual método de explicação você mais gostou? Por quê?

Do segundo, o LIME. Porque acho que visualmente é mais fácil, mais rápido de entender. É só bater o olho e entender o que está acontecendo que o outros dois. Você pensando no contexto de saúde pública, onde tempo é precioso. E é o que dá menos margens de erro para você interpretar. Pra mim é universal, qualquer pessoa consegue interpretar do que os outros dois.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Então, eu não sou contra. Mas eu acho que é uma coisa que deve ser aperfeiçoado daqui pra frente e eu acho que não substitui a experiencia de um profissional. Eu acho que pode ser uma coisa que ajude numa triagem bem inicial. Como por exemplo a análise da Citologia hoje, pode ser feito por técnicos e quando o técnico vê algo diferente ele chama o profissional médico para ele ver o diagnóstico. Eu acho que isso aqui pode ser tipo uma peneira mesmo dos casos mais prováveis e daí o médico confirmar se é isso ou não. E a partir daí passar a conduta. De repente em grandes serviços, onde tem deficit de profissionais ou lugares de difícil acesso, onde você tem que contar mais com a automação do que o profissional em si. De repente ajudaria de alguma forma. Nós

já fazemos isso de alguma maneira, talvez ajudaria muito para um médico generalista, alguém que esteja atendendo um paciente e não fosse um especialista, um enfermeiro, que realmente precise de uma atenção primaria e depois encaminhar para um especialista. Talvez seria de serventia.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Eu acho que pode ajudar e somar bastante. É viável. Só acho que tem que melhorar bastante e não só isso né, é necessário treinar os profissionais e ter um range grande de pessoas para pode testar cada vez mais o programa para ver se chega a algum consenso ali para poder liberar né.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Eu acho que de repente, você disse que usou dados de um hospital de Caracas né? Eu acho de repente você poderia testar com dados de outras unidades né, trabalhar com dados reais da população brasileira, ou até mesmo de outros países. Quanto mais troca você tiver, menor será o risco ne. Claro que existem riscos que são comuns a todas as populações né, mas tem riscos muito pertinentes a cada país. Por exemplo na África tem muito HIV positivo. Ali o maior fator de risco é ser imunossuprimido. Outros são países com alto nível de tabagismo ou menor acesso a saúde para prevenção de câncer, como exame de Papanicolau. Tem pais que tem maior cobertura vacinal para HPV, tem pais que tem menor cobertura. As particularidades é que temos que mudar né. Acho que é mais avaliar de uma forma geral e da a melhor forma de montar o programa de uma forma igualitária. Acho que é isso.

A. Entrevista PO06

Data: 03/02/2020

Onde: Consultório Particular / Gávea

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Aqui é 91%? O exame está aqui, mas esse resultado é sim ou não? Então com o exame, com o Teste de Schiller e independentemente dos resultados que você teve? Eu acho difícil explicar. Esse caso específico, eu entendi o resultado, mas não faz muito sentido esse resultado. Eu não sei como ele chegou a esse número, não é uma tendência que a gente ver na prática clínica. Vejo que é um caso fora da curva. Eu não sei como explicar, aí o que faz mais sentido pra mim... Essa idade em vermelho, jovem né, por que seria um fator de risco pra ela? Isso que você quer saber?

Eu consigo entender aqui os fatores de risco, mas para mim não faz sentido, com a tendência que a gente avalia e subjetivamente. Uma mulher jovem, sem Colposcopia, mas com Teste de Schiller, enfim. Mas eu consigo entender esse resultado aqui, só não consigo explicar como eles levaram, a esse resultado. Isso eu não consigo entender.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Bom, pelo que você a me apresentou aqui foi o Teste de Schiller, O Papanicolau e o Número de parceiros sexuais.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Não. Nesse caso ai não.

Acho que não é compatível com os dados de experiencia clínica. É uma justificativa que eu tenho para te dar. O fato de ela ter ou não o Exame de Papanicolau não tem significado, mas o resultado do exame, que não é isso que no modelo ele mostra. Essa idade precoce também não faz sentido pra mim como fator, bem, apesar de ela ser muito jovem. Para mim foi mais ter feito ou não o Papanicolau. Para mim faria mais sentido, qual o resultado do Papanicolau. E de alto grau ou baixo grau? O tipo de resultado, pelo menos agrupado. Agrupado com alto risco ou baixo risco para câncer.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada nestes métodos?

Eu acho que faria mais sentido aqui, uma citologia de alto risco e baixo risco para câncer. Você agrupar os resultados citológicos em baixo risco e alto risco. Porque isso vai dar um resultado favorecendo ou desfavorecendo, ter ou não Papanicolau não muda nada. E a mesma coisa é para o resultado da Colposcopia, uma colposcopia de alto grau e baixo grau. De baixo risco e alto risco para câncer, agrupando os resultados. Eu colocaria assim, ao invés de sim ou não. Isso faria muito mais sentido aí. Com as informações clínicas que a gente usa hoje. Para mim isso faria mais sentido.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

É o mesmo caso? Como eu explicaria? Pelo o que você esta me mostrando ai, ter o Teste de Schiller, não ter colposcopia e ter o Papanicolau, são fatores de risco para ela ter câncer. O que você está mostrando ai é isso. E não ter DSTs e não usar contraceptivos DIU são fatores para ela não ter câncer, mas não tão importantes como os fatores para câncer, eles tiveram um peso menor.

Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, a Colposcopia e no caso não ter DSTs foi um fator de proteção. Foram os três mais importantes, você juntando fator de risco e proteção, esses foram os três.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência

Artificial diagnosticou o câncer da paciente.".

Eu entendi o que ele usou para chegar, mas também ainda falta um pouco de sentido em relação a clínica. Porque é a mesma coisa que eu disse anteriormente, ter ou não a Colposcopia, ter ou não o Papanicolau, não seria um resultado que pra mim fazem sentido, para pesar se ele tem câncer realmente. Quer dizer, não ter Papanicolau para mim seria um fator de risco, mas até que no caso dela pode ser que isso tenha algum sentido. Acho que você pegou um caso fora da curva. Acho que você precisa verificar a tendência clínica, até para ajustar a sua coleta de informações.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Acho que aqui, eu não colocaria sem câncer e câncer, eu colocaria baixo risco e alto risco. Ou fator de risco e fator de proteção. Acho que ficaria mais claro para ver. E eu acrescentaria como dados polidos, não do método né, mas aí para qualquer método né, a mesma coisa do anterior. Ter qual tipo de resultado da Colposcopia e qual tipo de resultado do Papanicolau. Acho que são os fatores que podem mudar muito isso aí.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Esse aí já é mais difícil. Pelo o que eu estou vendo né, está em ordem decrescente né? Então o que pesou mais para o câncer, foi o Teste de Schiller, Idade e Número de gestações. É isso que ele está mostrando como fatores de riscos mais importantes.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente ?

Teste de Schiller, Idade e Número de gestações

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Ela é compreensível, mas para mim também falta sentido clínico. Até que a idade da 1º relação faz algum sentido, mas não acredito que seja o mais pesado. Mas de clareza eu consigo ver. Mas esses números são difíceis de interpretar para mim, não são tão claros. Eu posso comparar com os outros? O gráfico do 1º (SHAP) que você me mostrou me

parece muito mais claro. Aí eu vou ter que olhar o 00 e olhar aquele peso ali, para mim a diferença entre um e outro fica menos clara. Perde clareza para você ver os pesos de cada um, você tem que avaliar os números que são pequenos né.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Eu acho que o peso poderia ser apresentado graficamente do que numericamente. Eu acho que são números pequenos, ou senão um multiplicador que pudesse dar mais clareza nesses números. O primeiro me parece muito mais pesado que os outros, mas como é em números decimais, você não consegue visualizar com tanta clareza. Eu acho que isso mudaria. E esse + e – é desvio padrão? Isso também não ficou claro aí no gráfico não, só o peso. Não dar para saber o que é esse número do lado aí. E outra coisa aí, eu sei isso, porque eu e entendo um pouquinho de para o médico que não sabe estatística ou nunca estudou estatística, talvez não vá saber que isso é um desvio padrão. Então no gráfico a gente ver melhor, o gráfico é bem mais visual, é mais claro. Esse vermelho e azul, o tamanho da barra. No gráfico fica mais fácil, pelo tamanho da barra

Qual método de explicação você mais gostou? Por quê?

O primeiro, qual o nome? O da barra, O SHAP. Para mim é o que dá mais clareza. A visualização gráfica dele é mais fácil. Avaliar a importância de cada item e visualmente é bem mais claro de interpretar.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Sim. Porque acho que os dados, um número grande de dados eles tem uma representatividade, nos dá uma informação que talvez a gente não consiga visualizar sozinho. Acho que isoladamente não é suficiente, não basta. Mas eu acho que eles têm muito valor.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Sim. Porque eu acho que vai permitir chegar à conclusão, talvez o desenvolvimento de métodos menos invasivos, mais simples e talvez mais confiáveis.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Que sugestão? Eu acho que tem que ter um conhecimento clínico para escolher os fatores. Sem conhecimento clínico dos fatores é muito difícil desenhar, partindo dos

dados que já tem. Eu acho que é importante pensar na a fisiopatologia da doença para você conseguir buscar quais fatores que vão te ajudar nisso. Pelo menos no primeiro momento. Porque o que vi aí de crítica é que tem elementos e variáveis que para mim não fazem tanto sentido clínico, para justificar um risco de câncer. Talvez elas podem estar enviesadas aí por alguma coisa, porque os estudos clínicos não suportam esse achado que você teve aí. Ter ou não exame e qual o resultado do exame? Então eu acho que deve ter uma avaliação mesmo da clínica para chegar as variáveis que você precisa colher. Acho que pode ser utilizado como um tomador de decisão sim, só depende do que você vai usar como variável. Acho que pode, tem muitas variáveis que você pode utilizar para tomar decisão sim. Hoje já tem exames, exames muito específicos para definir o risco de câncer que podem ajudar na tomada de decisão sim.

A. Entrevista PO07

Data: 03/02/2020

Onde: Fiocruz / IOC

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Bem, se eu entendi, esse modelo está relacionado a diversos fatores, que quando acorrendo juntos, levou ao resultado do câncer. Por exempli aqui fala do Exame de Papanicolau, Teste de Schiller e número de parceiros sexuais. Esse conjunto de dados juntos, explica a probabilidade dessa mulher ter câncer de colo de útero.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O Teste de Schiller, exame de Papanicolau e o Número de parceiros sexuais. São três né? Então são esses três.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Por que ele teve câncer? Sim, por causa dos fatores de risco. Ele mostra aqui no gráfico né. Porque aqui ele mostra quais são os fatores que estão determinando essa condição da qual a influência deles para o diagnóstico. São fatores que são levados em consideração normalmente na clínica médica.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Visualmente eu não senti falta. Para mim isso aqui está bem claro para mim. Pra mim está bom. Eu não vejo dificuldade de entender fatores relacionados ao risco e o de menor risco. Talvez o estágio da doença, para poder ajudar o médico na terapia, não só para o diagnostico né. Ah ela está com câncer, mas qual é o estágio daquela doença? Qual o nível doença? Talvez aqui mostre que provavelmente ela tem câncer, mas e para o médico? Mas qual o estágio delas? Para poder direcionar um tratamento, para direcionar a próxima etapa. Então aqui está muito bem, mas eu acho que informação do estágio em que a paciente esta do câncer, tipo estágio 1, 2 e 3. Mostrar se essa paciente está num estágio mais avançado. Não sei se isso é possível. Mas eu acho que seria uma informação legal. Por que o médico pega aqui e agora? O que eu faço?

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutaion Importance?

Ele levou em consideração o Teste de Schiler, foi o que teve o maior peso pra ele chegar a essa predição. Ele também correlacionou a Idade da 1º relação sexual e o número de gestações. Agora é engraçado que ele colocou o Exame de Papanicolau ficou em quarto nesse método, então não considerou tão relevante. Achei bem estranho, o Papanicolau não ter tido um peso maior.

Estamos testando a clareza da explicação dada pelo Permutaion Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiler, Idade da 1º relação sexual e número de gestações.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutaion Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Não. Eu não consegui entender muito bem a relação da idade da 1º relação e número de gestações. O teste de Schiler sendo o primeiro de mais peso, faz sentido. Mas eu não consigo ver exatamente o quanto pesa a idade e número de gestações. Eu acho que a idade e o Exame de Papanicolau façam um peso maior, façam mais sentido.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Eu não sabia, sífilis não tem influência nenhuma? Eu só não consegui porque a idade e o exame de Papanicolau têm pouca influência. Ela tomava ante concepcional? Esse

valor é que? Ahh ta são todas as DSTs... Não, em relação ao método não. Mas para o trabalho geral, eu acho que aquela sugestão que te dei lá, seria interessante para qualquer um desses métodos.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Ele levou em consideração dois fatores de peso maior, que são o Teste de Schiller e Colposcopia e em terceiro o Papanicolau, mas com um peso bem menor do que a Colposcopia por exemplo.

Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Mais uma vez foi o foi o Teste de Schiller, que teve um peso maior para a decisão. Colposcopia e Exame de Papanicolau.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Sim. Porque além do teste tem dois exames complementares, que é a Colposcopia e o Exame de Papanicolau, que podem ajudar no diagnóstico. São exames que são usados na clínica. Para mim faz muito sentido. Acho que é isso.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Idade, se bem que ele colocou a idade aqui. Eu achei bem claro. Tem tudo.

Qual método de explicação você mais gostou? Por quê?

O que eu mais gostei? Gostei mais deste, o LIME. Depois, o primeiro que você me apresentou, o SHAP. O segundo que você me apresentou, foi o que menos gostei, o Permutation Importance, pois não consegui entender o que pesou nos fatores de risco.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Confio. Por que eu confio? Porque a taxa de assertividade ela é alta...e vem sendo provado ao longo do tempo. Obvio que precisa trabalhar mais essa parte que você vem explorando, que é a explicabilidade. É realmente como você falou, é uma caixa preta. O

que ele está fazendo é difícil de entender, mas a sua proposta é muito boa. Acho que é isso.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Sim, eu espero que sim. Porque eu acho que a decisão em conjunto de inteligência artificial com o médico, pode melhorar o tempo do paciente, proporcionar uma rapidez, melhoria de tempo. Se eu tenho um diagnostico baseado em IA, o tempo que aquele paciente ficaria fazendo um monte de exames, podem ser simplificados. Baseados nos resultados da IA, combinado com o médico, então ai o tempo de diagnostico já vai ser menor. O tempo de iniciar o tratamento de um paciente será mais rápido. IA auxiliando no processo terapêutico, predição de potenciais resistências ou não, eu acho que pode ajudar. Eu acho que tem que ter nessa parte, eu acho que não só nesse sentido. Mas também no corpo médico, administração de hospital, financeiro, no hospital como um todo. Acho que tudo pode auxiliar e ter mais qualidade e eu espero que mais assertividade também. Assim eu espero.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

A validação do processo, é isso que você está fazendo né. O médico pelo o que eu tenho convívio, gosta de ter uma certeza mesmo. Então se você levar pra ele todos os argumentos, provar e fazer um teste, eu acho que eles podem aceitar bem. Aqui você não está excluindo-o, aqui você está somando. A informação que eu também acho muito legal, é mostrar o estágio desse tumor, eu não sei hoje, mas futuramente você incluir coisas novas para o médico. Acho que a probabilidade de ajudar é muito grande, mas para isso é importante mostrar uma clareza e uma certeza para ele sobre o método. Eu acho que é isso. Eu não sei o que mais eu posso falar. Eu usaria, mas não sei outros médicos. .

A. Entrevista PO08

Data: 10/02/2020

Onde: Fiocruz / IFF

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Eu diria que O resultado de predição de câncer está influenciando pelo fato dela ter realizado o teste de Schiller, colposcopia, exame de Papanicolau. Ter o número de sexual de parceiro para mim não tá claro. Número de parceiros, se é ter essa resposta ou ter um determinado número, a idade é a mesma coisa né. Olhando assim, é porque eu já vi nos outros métodos que era ter realizado Schiller ou ter realizado colposcopia né. Mas eu acredito que o número de parceiros, depende do número absoluto de parceiros e a idade e a idade também né. Olhando aqui, essa tabela complementa essa informação né, aqui Colposcopia e Schiller é 1 ou 0 né, as doenças também é 1 ou 0. Mas os números de parceiros é o que tem um número absoluto, 4 e 21. Então assim, essa tabelinha complementa esse gráfico. Eu diria que fica mais inteligível.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Isso não está tão claro. Tem essa barrinha né. Essa barrinha está maior no Teste de Schiller e na Colposcopia do que nos outros né. Então acredito que seja isso, intuitivamente me parece que essas duas variáveis foram as que mais influenciaram.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Concordo, mas não completamente. Porque você precisou me explicar. E Olhando o gráfico e a tabela, eu tive que deduzir a partir da explicação inicial e estou olhando bastante para deduzir e tentando entender. Eu diria que nunca vi um resultado desse, é bem interessante. Torna visível coisas que eram invisíveis até então. Então tem bastante utilidade pra você ser convencido da predição que ele faz né. Faz sentido que essas variáveis sejam as mais influentes e faz sentido que elas influenciam na predição da presença de doença.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Considerando o quem tem na base de dados, que só se realizou ou não né. Eu sinto falta do resultado de alguns exames. Então por exemplo, o exame de Papanicolau ter feito ou não, como exemplo toda mulher entre 25 e 65 anos tem que fazer o exame de Papanicolau. Ter feito ou não, eu acho que isso me influencia pouco, tanto é que aparece como pouca influência ai né, não dar nem pra ver a barrinha, a barrinha é muito pequenininha. Agora, o resultado do exame faria mais sentido pra mim, dado que o resultado do exame aponta para uma doença mais grave, isso influencia na predição de presença de doença.

Mas ai deveria ser contemplado no próprio algoritmo né, pois o algoritmo foi construído pelo fato de ela ter feito ou não aquele exame né. Então eu como técnico, pensando na aplicação do algoritmo eu acho que ele melhoraria o desempenho se ELE tivesse essa informação. Ai em consequência, assim como está escrito ali, Exame de Papanicolau, eu colocaria Exame de Papanicolau com a doença X. Tornaria mais claro porque o modelo fez corretamente a predição. Agora com o que você tem na base, eu não vejo como você tornaria mais claro. Eu só vejo como melhorar o algoritmo. Agora tornar mais claro, eu não vejo, sinceramente não está me ocorrendo. Apesar de eu nunca ter visto, você me apresentou outros dois modelos. Esse ai está deixando mais claro para o usuário, acreditar porque ele está acertando nessa predição. Então assim, resumindo, eu não saberia te dizer alguma coisa que eu estaria sentindo falta. Seria mais melhoria do algoritmo propriamente dito.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

O que mais influenciou a previsão da presença de doença são as variáveis que estão em vermelho, isso visualmente é bem interessante. Mas não aparecem todas as variáveis, só aparecem parte delas, então o outro (LIME) tem a vantagem de mostrar todas as variáveis

que influenciaram ou deixaram de influenciar na predição. Então me parece que o outro (LIME) me dar mais transparência, explica melhor.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Novamente Teste de Schiller e Exame de Papanicolau. Que estão com a barrinha maior. A terceiria seria a seguinte que é. Na verdade a terceira não está claro pra mim, se é o Números de parceiros sexuais ou não apareceu na etiqueta né.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Sim. Depois de ser apresentado a figura, mostrar o output do modelo sim. Concordo. O diagrama de cores e o tamanho da influência de cada variável.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Que ai não aparece todas as variáveis usadas na predição. Ai aparecem pelo visto as mais importantes, tanto para predição de doença, quanto para a ausência delas. No modelo anterior (LIME) , lista todas elas. A vantagem do outro é que mostra todas as variáveis, ai fica claro dever o que está influenciando e o que não está.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Impraticável. Estou tentando entender o output. O peso e a variável. Está em ordem? Está vendo, não consigo nem ver se ele está em ordem. Visualmente ele é muito ruim, a informação está ai, mas eu tenho muita dificuldade de entender. É desvio padrão? Está vendo, eu não consigo nem entender o que é a medida. A primeira medida eu sei que é peso, a segunda medida é desvio padrão ou é faixa? Eu não sei né, não está escrito. A primeira coisa de uma tabela é que ela precisa ser auto explicativa né, e você não vê isso no output. Sendo faixa ou desvio padrão, aqui parece que mais influenciou foi Teste de Schiler, Idade da relação sexual e número de gestações. Que não é o mesmo dos outros modelos, Como é que um modelo de análise de desempenho de um algoritmo pode trazer informações diferentes? Isso me levantou uma dúvida né.

A capacidade de explicação de um modelo é dada pelas variáveis que estão lá e o algoritmo é o mesmo, como é que uma ferramenta de tentativa de explicação de um algoritmo que dá resultados diferente de outra? E como a gente vai medir o desempenho da explicação ne? Aquilo que você estava me falando, não tem né? Você só tem o desempenho do modelo preditivo, mas não do modelo de explicação né. E isso está me deixando em dúvida, em qual modelo eu acredito ne. Eu estou respondendo quais que tem mais informação e as informações mais inteligíveis ne, mas não sei qual o mais confiável, já que trazem informações diferentes.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, Idade da relação sexual e número de gestações. Mas isso depois de perguntar muito pra você, porque estava difícil de entender.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

O quanto? Pouco. Porque eu tive dificuldade de entender a tabela de saída.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Eu senti falta de alguma nota de rodapé da tabela. O que é o peso, que medidas são essas. Coisas que nos outros modelos que tem uma diagramação visual, fica muito mais fácil de entender. Nesse ai, eu tenho que ficar olhando os números e apesar de facilitar vir em ordem de grandeza e decrescente, mas não é tão fácil quanto nos outros não. Nos outros eu tenho esses pesos distribuído de forma visual e em cores. Os outros são muito mais claros em termos de explicação.

Qual método de explicação você mais gostou? Por quê?

Gostei mais do LIME. Porque ele é claro e é visualmente completo. O SHAP eu também achei claro, mas ele não tem todas as variáveis. Esse aqui num único output, eu tenho a probabilidade da ocorrência de doença, a probabilidade de não ter doença, tenho a lista das variáveis, tenho a distribuição dela influenciando num sentido ou em outro e tenho o peso de cada uma delas. Então eu acho ele mais completo, foi o que mais me interessou.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Sim. Porque eu acho que ele trabalha como a cabeça da gente trabalha ne. Só que a gente não faz cálculo né, mas a gente sempre raciocina probabilisticamente. O que a IA faz é usar essas informações, atribuir valores e pesos, para chegar em uma conclusão né. Então eu acho que são uteis e serão cada vez mais uteis.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Com certeza. Deixando para o domínio médico as tarefas de maior complexidade que envolvem tomada de decisão, condutas, procedimentos.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Olha, eu acho que ele seria muito bem utilizado. Não sei se eu estou entendendo a sua pergunta. Você está falando em termo de aplicação né? Eu acho que a sua popularização. Não só que eles tenham sua acurácia medida e divulgada, que eles sejam popularizados para que a gente possa utiliza-los na pratica clínica. Seja por exemplo, o próprio usuário do sistema de saúde avaliando pelos seus sinais e sintomas, se precisa procurar o médico ou não. Seja por um setor de triagem que precisa se um paciente precisa de atendimento mais rápido ou não num sistema de agendamento do que outro. Hoje por exemplo, para podermos agendar uma paciente com algumas lesões, depende de algum profissional regulador, e esse profissional se ele trabalhar muito bem, ele faz direito. Mas o que a gente vê na pratica são muitos erros de avaliação, colocando todos os paciente no mesmo nível de prioridade, resultando em filas incoerentes, onde você tem pacientes com maior gravidade sendo atendido depois de pessoas com menos gravidade. Então eu acho que um sistema de inteligência artificial desses seria muito útil nisso. Quando eu falo popularização, eu também falo em redução de custos.

A. Entrevista PO09

Data: 11/02/2020

Onde: Fiocruz / IFF

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

O que entendi foi o seguinte. A parte gráfica né, o lado direito da parte gráfica temos fatores que contribuíram e o seu respectivo fator de contribuição para o câncer, e do lado esquerdo os fatores que falam negativamente para o câncer, ou seja, aquele fator fala para câncer. Na tabela o que aparece em azul fala contra e o laranja é o que fala a favor. Nesse eu consigo entender o gráfico e a explicação, eu só não concordo com os parâmetros. hahahaha

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Esse Lime? Teste de Schiller, Colposcopia e Exame de Papanicolau.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Concordo parcialmente. Porque assim, apesar de eu conseguir entender qual foi o fator que ela usou, eu não sei qual o racional por trás disso. Alguns desses fatores eu acho que estão estranhos, como eu não sei o racional por trás disso, ele não responde a minha questão principal. Por que ele valora isso como positivo ou não? Eu sei que ele tem um algoritmo e uma variável de dados. Eu sei que tem uma modelagem matemática por trás disso, que eu desconheço. Mas eu não sei por que ele escolheu aquele padrão ou aquele

perfil para colocar. E essa não está explicado ali. Assim, ele diz que utilizou aquilo, mas ele não diz como utilizou. Por exemplo o Teste de Schiller? É sim ou não? É positivo ou negativo? Não diz qual foi o dado do Schiller para ele avaliar. Foi o Schiller, mas foi ele fez ou não fez? Positivo ou negativo? Isso não tem, então isso faz diferença na hora de eu validar e entender se a premissa dele está correta ou não. Se o raciocínio dele está correto ou não. Grosseiramente falando assim, algumas coisas eu consigo entender e outras não. Eu sei qual foram os parâmetros que ele usou, mas eu não sei como ele usou esse parâmetro. Por isso que eu concordo parcialmente.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Que tenha acesso no parâmetro para abrir e descobrir como funciona. Ahh, eu quero saber como ele utilizou o Schiller, daí ele diz foi utilizado isso, foi considerado aquilo. Mostrar qual o parâmetro de base utilizado no Schiller. Se eu achei estranho porque ele utilizou isso ou aquilo. Queria poder ver o racional por trás disso, quando eu clicar no nome. Porque ele usou isso daquela maneira. Qual foi o racional por trás daquilo pra dizer se é importante ou não. Eu sinto falta disso. Provavelmente todos os métodos terão o mesmo defeito. Eles todos vão me dizer qual que tem influência positiva ou negativa, e todos vão me falhar em explicar o porquê. Então ele não explica, só informa.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

De novo ele acha que o Schiller é o fator mais importante, seguido pela idade da primeira relação sexual e número de gestações. Depois número de parceiros e o que aparece como negativo ficam abaixo, como as DSTs e embaixo mais 15 que não estão aparecendo. Aqui você tem que entender um pouco de estatística, porque ele não diz o que é esse valor, ele só te dar uma média e o erro padrão. Eu até entendo que ele colocou esse primeiro que aquele, pois esse tem o erro padrão menor. Que parâmetro é esse? É média? É desvio padrão? Mas eu não sei que valores são esses. Eu entendo que ele considerou esses valores como mais importantes, mas porque eu também não sei. Eu não sei que cálculos são esses. Não sei.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiller, idade da primeira relação sexual e número de gestações.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Não. Que eu entendesse não. Ele me mostrou quais foram os parâmetros que ele utilizou. Mas a explicação do porquê não está aqui. Ele me informa os parâmetros que ele utilizou, mas a relação que ele faz entre o parâmetro e o câncer eu não sei. Isso ele não explicou. Porque olha só, ele me diz aqui Teste de Schiller, idade da 1º relação, número de gestações. Mas percebe que nem o resultado é igual ao outro (LIME). Ele te dar outras informações. Que variável é essa? Eu não sei que conta é essa, se é risco, se é risco relativo, eu não que conta é essa.

Eu percebo assim que ele categorizou as variáveis pra chegar a decisão, mas ele não me conta como ele pensa para chegar a decisão. Porque o Teste de Schiller? Como ele utilizou? O mesmo defeito que o LIME tem, esse tem. Ele não me mostra o racional por trás daquilo. Ele me mostra um número e me mostra de maneira seca, olha esses são os parâmetros mais relevantes. Se é uma plataforma que você está acostumado a usar e te diz como o Schiller é trabalhado, por exemplo a idade da 1º relação foi jovem. Esse aqui eu sei o racional por trás disso, já o Teste de Schiller eu não sei como foi utilizado. Se é positivo ou negativo? Se fez ou não fez? Não está escrito, então eu não sei, baseado em que? Positivo ou negativo eu consigo entender, se foi feito eu não já não consigo entender. Então assim, eu não sei o racional por trás disso. Por exemplo o número de gestações, só teve uma, isso é ruim? Eu não sei como ele utiliza esse dado. Então tem alguns dados ai que eu não consigo entender. Ahh, o Número de parceiros sexuais eu entendo, quanto mais você se expõe, mais risco de adquirir HPV que é um fator de risco principal.

Então tem coisas que mudaria, tem coisas que seriam fundamentais para esse algoritmo que não estão aqui. Minha compreensão de fator de risco. Mas é um dado que você não tem, então não tem como colocar. Tem dados aqui que tem lá sua contribuição, mas que não são tão relevantes. Por exemplo, a idade ter contribuído como quinto fator de risco. Ela com 21 anos? 21 anos não é fator de risco para câncer de colo de útero. Está fora da faixa. Então, por que ele usou a faixa etária dessa forma? Mesmo assim, você ver que é baixo, muito menos que 1. Eu não sei que índice é esse. Então assim, ele não me explica, somente me informa. Informar é uma coisa, explicar é outra totalmente diferente. Essa é minha opinião.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Bem, eu senti falta da mesma coisa. Do racional por trás disso? O que tem nessa base? O que é esse número? Isso aqui eu entendo que é desvio padrão. Mais ou menos eu entendo que é desvio padrão. Que é esse número eu não sei. Quando eu não sei o número, eu não consigo criticar. Então quanto mais perto de 1 mais importante? Quanto menos perto de 1 menos importante? Ou isso não? Tem algum algoritmo interno que nunca vou entender o que ele é. Não tem uma coisa aqui dizendo qual é o cott off ? Eu tenho uma sequência de números, qual é o cottof? O quanto isso é importante e a partir de quanto isso é importante? Ele viu um fator de risco, mas o número que tem ali é pequeno, então eu não sei. Porque o outro não me dá número, então eu não consigo entender. Esse da número, então quando da número eu quero entender. Não é tão simples.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Deu 91% de predição né. Bom, onde está o output value é o valor de percentual que ele diz que é o risco. É como se ele descontasse do valor principal, a Colposcopia o número de gestações e os outros valores negativos. É como se ele somasse os valores positivos do outro lado esquerdo. Ele diz que o Teste de Schiler, o exame de Papanicolau são mais importantes e o número de parceiro sexuais, a idade e outros fatores menores.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiler, o Exame de Papanicolau e o Número de parceiros sexuais.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Novamente. Ele mostrou as contribuições de uma maneira mais gráfica que os outros. Mas ele não explica, ele só me diz os parâmetros que ele utilizou usou e em que percentual ele utilizou esses parâmetros. Visualmente ele mostra, ele bota uma régua de valores negativos. Mas visualmente ele mostra. A impressão que eu tenho é que ele subtrai os valores negativos dos positivos e chega um valor. Mas ele não explica por que essas variáveis estão ali. Ele me mostra baseado em que ele tomou essa decisão. O quanto pesou essa decisão para chegar em 0.91%.

Não em forma de grandeza, mas de outra forma. Ele mostra uma régua, como uma escala né é de 0 a 1 e pouco né? Enfim, então eu não sei se entendo isso como 91%, eu

não sei ate onde vai essa escala. Mas, novamente ele tem os mesmos erros dos outros. Mas visualmente ele é mais fácil de entender do que os outros. Se bem que os outros não são défices, mas visualmente esse é o que melhor representa. Mas ele não explica novamente.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Nesse aqui eu não sinto falta de elementos visuais. Eu sinto falta de um link ou de uma explicação para entender como cada um desses fatores. Como esse dado entra? Como ele foi tratado? É positivo ou negativo? É contínuo ou não é contínuo, esse banco né? Ou como é que ele surge? Por que ele pensa a premissa dessa forma? Quais são os dados de conhecimento prévio desse banco? O que faz com que eles valorizem esses dados?

Qual método de explicação você mais gostou? Por quê?

Visualmente eu achei melhor o SHAP, em seguida o LIME e por último o Pemutation Importance. Porque eu sou patologista e lido com imagens. E essa imagem se auto explica pra mim. É intuitiva essa imagem. A LIME também não e diferente, é uma questão de gosto né. O outro (Permutation Importance), como é uma tabela e é número. Número sem explicação pra mim me incomoda, eu preciso saber o que quer me dizer. Ele não me diz, só me mostra os números, todos menores que 1. Então ei não consigo avaliar, porque eu não sei se é uma análise de relação, se é uma análise ade risco, entendeu? Ele da um valor que eu não sei se é risco, então eu não consigo entender o dado.

Então o porquê, nenhum deles me explica. Ele só me diz o seguinte, os parâmetros que eu valorizei mais são esses, seguido desses e tal. Mas o porquê de ele chegar a esse parâmetro ele não me diz. Então como eu sou uma pessoa que quando quero saber o porquê e o racional por trás, ele só me diz os parâmetros. E claro que se você já conhece o método, você vai saber os parâmetros e consegue entender o que ele usa do banco, se é sim ou não ou positivo ou negativo. E claro que tem uma modelagem matemática que você não vai entender, mas você tem que entender por que do banco de dados aponta para aquilo

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Sim. Eu acho que se for feito da maneira adequada, acho que tem tudo para funcionar. Eu não vejo por que não. Eu só acho que é complicado você fazer a migração direta, enquanto você não tem uso. Para você confiar numa tecnologia, você tem que usá-la e

comparar com o que você sabe hoje, até pra provar que ela é melhor. Eu acredito que ela vai ser melhor, porque a gente não tem condição de usar o cérebro para fazer big data. Então mesmo que não seja sozinha, a IA já ajuda em muita coisa. Então acho que é uma coisa que tende a ter um potencial absurda de aplicabilidade e de facilitação.

Só que o que eu temo na IA é que quando eu não controlo como esses dados são gerados, fica difícil para eu criticar. Então você vai acabar ficando refém de uma tecnologia que se ela dá errado, você vai prejudicar muita gente. Se você não tiver um parâmetro muito claro para entender, o porquê e onde a máquina pode errar, você vai se lascar. A maioria das máquinas que eu uso, eu sei os fatores críticos. Então o fator crítico eu tenho que olhar. O fator crítico é o limitador. Eu também tenho fator crítico, o ser humano tem essas questões. Quando eu chego no meu limite eu tenho que checar e rechecar. Então quando eu vou rechecar a máquina? Ou vou aceitar tudo passivamente? Eu sinto falta dessas questões. Se você não tem isso muito claro, não que o método tenha que me informar isso. Mas se você não tem muito claro quando usa uma tecnologia, a limitação dessa tecnologia, acaba sendo um desserviço. Agora se você conhece todos os riscos e benefícios e sabe a melhor forma de aplicá-la, eu não vejo problema nenhum de utilizá-la.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Pode. Acho que pode. Porque o diagnóstico nada mais é do que você processar dados e chegar a uma conclusão. A questão é, tem coisas que você nunca vai substituir. Algumas coletas de dados depende de quem faz, da mesma forma que se o médico não sabe coletar informações clinicas. Se você alimentar o seu banco de dados errado, você não vai funcionar.

Então cada metodologia tem a sua limitação, mesmo a IA não vai ter tudo. Então assim o fator humano é, nem tudo a pessoa vai responder de verdade para a máquina, pode omitir mais ou menos, depende do quanto ela confia na questão da segurança dos dados dela. Tem coisas que você conta para um médico e não conta pra outro. O médico pode não valorizar outro vai valorizar, então vai depender do seu background. É claro que você tem uma central enorme, com um banco de dados do mundo inteiro, com todos os desfechos muito bem documentado, a tendência é que você consiga fazer melhores decisões baseado nisso. Coisas que o cérebro humano não vai conseguir acessar, ainda mais como é um sistema que aprende consigo mesmo, a tendência é que ele vá se refinando. Mas isso não impede que eu tenha problemas, se eu tenho um banco de dados originalmente errado eu vou ter problema e vou ter viés. Se eu tenho uma nova variável que eu não tinha antes e não conhecia, muita coisa nova, como é que se entra nesse contexto? Então assim, tem

coisas que a máquina tem condição de avaliar que eu não consigo avaliar. Então eu acho que tem espaço. Eu acho que vai mudar.

Eu acho que muda muito na questão de avaliação de prognostico da paciente. Na oncologia cada vez mais dados, por exemplo, para o paciente saber a sobrevida em quantos anos. A máquina vai calcular isso melhor do que eu. A máquina vai poder te dar isso de maneira muito mais precisa do que eu, e ela pode errar, porque a estatística é previsão, não é realidade. Então quando eu lido com estatística, o médico lida com estatística, o computador lida com estatística, mas eu tenho que adaptar para o indivíduo. O computador pega estatística e faz uma leitura, baseado nos dados que você me deu. Se o seu dado estiver errado eu posso errar. Da mesma forma que eu não tiver todos os dados e informação, eu também posso errar. Eu vou valorizar coisas diferentes e isso é natural. Acontece na própria prática médica e vai acontecer com IA também. É só você alimentar o exame errado ali, você lascou o algoritmo inteiro. Então dependente da onde você pegou seus dados, da qualidade dos dados, de onde ele vem, você pode ter um bias importante de uma paciente. Então eu não acho que seja tão simples, não vai ser tão rápido de aplicar. Eu acho que vai demorar para a gente aplicar, mas eu acho que vai chegar. Não sei se eu vou chegar a ver no meu tempo de vida profissional útil. Mas eu acho que tudo implica para big data, a gente está na fase das ominicas. Isso vai chegar, vai baixar o preço, você vai conseguir fazer a avaliação de sequenciamento para todo mundo e você não vai conseguir avaliar isso só com tabelinha como a gente faz hoje.

Você vai ter alguma, nem que seja para fazer probabilidade de prognóstico para doente. Qual a resposta de drogas ele vai ter melhor. Então mesmo que o médico não seja substituído, que eu acredito que não seja, porque a relação humana é a mais importante no tratamento, até mais do que o conhecimento. Porque a adesão do paciente ao tratamento, não e só o dado bruto, até porque o dado bruto erra...então o médico terá até que se humanizar mais. Ahh o dado bruto é esse, mas ele erra ne. Então as vezes o paciente não interpreta legal um dado bruto e ele vai ter um surto ali.

Então assim. Tem coisas que eu acho que não vai substituir. O relacionamento humano nunca vai ser substituído pela máquina. A gente vê o problema do excesso do uso de máquina, na questão do desenvolvimento para as crianças. Porque tem umas coisas que ajudam melhor, mas se não souber fazer você vai perder em outras questões. Então eu acredito que o médico continuará tendo espaço, continuará tendo função, até porque você não vai processar uma máquina? Ele vai continuar prescrevendo, por questões legais. Mas ele vai estar armado de uma tecnologia que vai melhorar a forma dele abordar esse paciente. Aí eu vejo isso com bons olhos, só que eu acho que tem que ser muito bem feito, porque não vai ser assim, peguei um programa. Vai ter um monte de estudos,

validação, vai testar um monte de banco, vai ver se na vida real é igual ao que está no simulado lá. Até que um dia esse sistema pode entrar, como já entrou outros. Os sistemas de inteligência artificial já ajuda no diagnóstico de um monte de coisas. Então assim, eu acho que essas coisas vão entrando por áreas, por aplicações especificas, até que um dia você terá um super big data ai.

E que vai conseguir fazer um monte de diagnóstico e talvez a gente resolve o problema do diagnostico ou dos prognósticos, ou dos fatores preditivos de respostas a mais drogas. Eu acho que aplicabilidade para isso. Mas também se não tiver aplicabilidade, de reduzir custos e aumento de acerto, a ferramenta não se sustenta. Então ela tem que ou diminuir custo, ou diminuir prazo ou aumentar o acerto. Uma dessas coisas ela tem que fazer para ser útil, até porque não será uma coisa barata de implantar né. O desenvolvimento disso deve ser caro pra caramba, pra você fazer isso funcionar.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Eu acho que o racional é a principal coisa que tá faltando. Não serve só o como você usou. Mas como esse dado é alimentado e como ele é valorado. Porque que eu entendo que essa paciente com 21 anos tem risco de tantos por cento de câncer? Ah, não é som porque ela fez Papanicolau, o Papanicolau dela deu lesão atípica, deu alteração? Tem alguma coisa a mais que eu não estou vendo, eu só estou vendo como se fosse uma tabela e as variáveis que eu estão trabalhando. Não me diz como usa essas variáveis. Qual a variável que puxa pra cima e a variável que puxa pra baixo? Para eu entender esse cálculo, isso eu não tenho. Eu não sei como o programa tratou essas variáveis. Ou ele descobriu uma modelagem nova que ninguém nem sabia a relação? Então eu não sei.

Eu sinto falta da explicação do racional da base. Ele não me explica porque esses parâmetros. Ele não explica porque escolheu esses parâmetros. E porque era o que eu tinha? Ah, então tudo bem. Mas quem coletou esses parâmetros, coletou de que forma? Que forma colocou esses dados? Pra eu poder indicar esse tipo de tratamento no meu dado, eu tenho que entender o racional por trás disso. Até para poder ler o resultado. Então assim, para quem sabe como os dados foram tratados, quem sabe como o computador avalia essa modelagem, não precisa de mais informações das que estão aqui. Para quem não sabe, isso não me satisfaz enquanto profissional da área. Eu queria saber mais.

A. Entrevista PO10

Data: 12/02/2020

Onde: Fiocruz / IOC

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Eu explicaria que o Número de parceiros sexuais, Papanicolau e Teste de Schiller foram os preditores que mais influenciaram para essa paciente ter câncer.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Número de parceiros sexuais, Papanicolau e Teste de Schiller.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Mais ou menos. Na verdade dizer que o Teste de Schiller, Exame de Papanicolau e etc. Não é uma explicação pra mim. Ele não explica o porquê. Então não explica, ele justifica.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

A explicação do porque esses parâmetros são bom preditores. Sei que essa informação precisa estar contemplada né.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo

sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Basicamente da mesma forma do outros. Ele da uma lista dos parâmetros e mostram os valores de cada parâmetro. Mas ainda não tem a explicação do porque cada um desses preditores chegaram a esse resultado. A única coisa que vejo de diferente é o layout, só mostra uma lista mesmo.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente ?

Teste de Schiller, Idade da 1º relação e número de gestações.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

De novo igual a outra. Ele justifica, mas pra mim não é uma explicação. Ele justificou os parâmetros que foram mais importantes, mas não detalha e explica.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutaion Importance? Que informação seria interessante ser apresentada neste método?

Mostrar um diagrama com o significado e detalhe de cada parâmetro. Um diagrama do que muda em cada parâmetro. Isso sim seria uma explicação, mas até agora ele não explicam. São modelos que justificam.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Basicamente uma outra forma visual de mostrar os mesmos resultados. Teste Schiller, Exame de Papanicolau e número de parceiros sexuais foram os preditores que mais contribuíram. Mas basicamente é só uma forma diferente de mostrar o resultado.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste Schiller, Exame de Papanicolau e número de parceiros sexuais

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência

Artificial diagnosticou o câncer da paciente.".

Sim. Ele ajuda a compreender. Mas sempre justifica, nunca explica. Essa representação gráfica eu acho mais complicado para ampla maioria das pessoas (leigos).

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

A explicação propriamente dita. Isso pra mim é uma justificativa. Era necessário um diagrama com detalhes de cada exame, o que está diferente em cada parâmetro desse.

Qual método de explicação você mais gostou? Por quê?

O SHAP em primeiro. Pela informação gráfica. Ser o mais bonito. Mas eu acredito que para leigos esse é mais difícil.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

A princípio sim. Tem áreas que está mais avançada. Na medicina vai envolver muita coisa relacionado a IA e ML. Nós confiamos, mas a princípio é testar e fazer estudos de como implantar os métodos. Mas eu tendo a confiar e acreditar. Eu confio nos resultados gerados. Mas não confio na ética...

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Sem dúvida. Pela velocidade do diagnóstico. Só isso já seria uma revolução. Aumentar o número de atendimentos, tanto em área urbanas que não tem médicos suficientes e também em lugares remotos. Só falando em agilidade já seria uma revolução. Fora maior exatidão, padronização, uniformização.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Na forma como estão hoje eu não vejo um grande benefício para o paciente. Não consigo visualizar uma melhoria direta para o paciente. Seria bastante interessante ter detalhe de cada exame né. O porquê de cada parâmetro desse aumentar a probabilidade de câncer da paciente.

A. Entrevista PO11

Data: 13/02/2020

Onde: Hospital OncoDor – Barra da Tijuca

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance ?

Acho que o que deve ter acontecido é que o algoritmo deve ter pegado os fatores que são mais frequentes nos pacientes que tem diagnóstico de câncer de colo de útero né. Acho que são os mais comuns entre os pacientes e ficou maior e mais importância, na verdade. Que significa que o Teste de Schiller, é um teste que quando alteram então significa que tem alguma alteração celular no colo de útero e que sugere que momento que a gente chega aí tem que levar alterada ele tem que seguir com a investigação Clínica. E independente dos outros fatores de risco, ele realmente é um dos fatores mais importantes. Eu não sei só a Colposcopia, que é uma exame importante, porque ela ficou mais abaixo? Mas talvez porque talvez possa ter algum viés ai nos dados né. Talvez nem todo mundo tenha feito a Colposcopia, eu não sei o que aconteceu nos dados. A primeira relação sexual tem lógica, porque normalmente sugere que o paciente tem mais tempo exposto ao vírus da HPV e aí tem mais tempo de evoluir com uma lesão pré- maligna para uma lesão maligna né, para o câncer. Então eu acho que é isso, acho que o algoritmos pegou os fatores de riscos mais frequências que tinham naquela população e que batiam com os pacientes com câncer de colo uterino.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Em relação ao modelo? Parece que foi o teste de Schiller, está em ordem né?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo

modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Sim. Sim. Porque na verdade ele pegou um exame que provavelmente é positivo, que é o Teste de Schiller e ai o paciente com o teste SCHILLER tem mais chance, porque normalmente é um exame pré diagnostico né. E aí conseguiu avaliar, por isso, conseguiu definir melhor. E os outros fatores são fatores de riscos que também aparecendo em ordem, número de parceiros ser maior em relação sexual e ai a pessoa é mais exposta ao vírus do HPV né e quem tem esses fatores de risco, maior chance de ter Câncer de colo de útero.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo? Que informação seria interessante ser apresentada neste método?

Na verdade ele tem o peso de importância e a ordem de importância já né e tem a ordem. Eu acho, que eu acho que ele é bem explicativo, talvez faltaria botar o que não ajudou, talvez dados que não ajudem? Mas ele coloca os em ordem de importância, então tipo assim, acho é bem explicativo na verdade.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

Eu achei esse mais interessante que o outro, um pouco em relação a coisa ele coloca a visualização. É melhor de você ver a importância das características, esse visualiza melhor. Lá (Permutation Importance) é só o número e esse você tem o gráfico, números e então acho que você visualiza melhor. E ele coloca também a mesma coisa que achei legal, como o Teste de Schiller como o mais importante né, e o exame do Papanicolau também que realmente é importante, porque é o exame aqui que altera antes do teste de Schiller. Então são realmente fatores importantes que ele colocou: o número de parceiros, idade da primeira relação sexual, enfim. Ele realmente define o que que tem mais poder e o que tem o menor poder de um algoritmo né. Então acho que esse método é bem interessante por conta disso, que ele visualiza, você consegue visualizar bem.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência

Artificial diagnosticou o câncer da paciente.".

Sim. Sim. Só não entendi uma coisa. Posso falar? Por que que a Colposcopia ficou como baixo o valor? Ahh ta, entendi. Ficou como não. Então ele considera a Colposcopia importante? É isso? Na verdade, ajudou pouco, porque ela não fez o exame. Como não considera importante. Achei esse método mais interessante para explicar.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Não vejo nada que falte na verdade. Acho que tem tudo. Nada acrescentar. É só explicativo né, talvez agora olhando esse, eu acho que realmente o outro (Permutation Importance) falta você ter essa visualização gráfica né.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Ele é bem parecido com o último né (SHAP), enfim, em relação aos critérios que colocou de importância né. Só que aqui ele colocou um pouco diferente, porque ele colocou DSTs aqui, que no outro vocês colocaram mais abaixo. Mas eles colocaram como importante, realmente o Teste de Schiller e colposcopia né. E ele também é bem interessante porque ele coloca no gráfico os fatores, enfim, igual o segundo né? Só que de uma forma diferente, enfim, mas a visualização dá para visualizar bem. Mas eu acho que no segundo (SHAP) ele explica um pouco melhor do que esse terceiro talvez.

Estamos testando a clareza da explicação dada pelo LIME. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Acho que foi o Teste de Schiler né. A colposcopia ajudaria, mas a paciente não tinha a colposcopia no exame. Mas eu acho que se tivesse, seria mais um fator que seria importante para o diagnóstico aumentar o poder estatístico. Mais um fator de risco para diagnostico do câncer. E ai ele colocou, na verdade como outros fatores o paciente tem histórico de DST e isso é uma coisa que é interessante, porque na verdade o paciente que tem DST acabam tendo múltiplos parceiros e não usam preservativos. Então poderia contribuir, aquela paciente ter uma predisposição maior ao vírus do HPV. Entendeu? Então ele explica, talvez o paciente ter o diagnóstico, por isso. Porque talvez seja um paciente um pouco mais promíscuo, no sentido de ter mais relações sexuais em sem preservativo. E o mais importante são os dois métodos de diagnóstico que são o Teste de Schiller e a

Colposcopia, que tem um impacto no diagnóstico do câncer de colo de útero.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente."

Sim. Sim. Porque ele coloca quais são os dados clínicos que a paciente tem que influenciaram o diagnóstico do câncer. Então os dois fatores de risco e ele deu importância a esses dois métodos de diagnostico.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Qual método de explicação você mais gostou? Por quê?

Eu acho que os dois melhores, são o segundo (SHAP) e o terceiro (LIME). Mas eu achei o terceiro melhor, eu acho que eu tenho mais dados ele te tá mais dados, ele dá o dado de sem câncer, e ele dá todos os fatores que influenciaram. E aqui você consegue ver todos os fatores importantes e os fatores que também não são importantes né. Mas acho que os dois últimos são os melhores.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

Tenho, eu acho que confio. O principal é você saber selecionar todos os dados né. Porque se você não tiver os dados certos, você pode acabar tendo um viés que acaba interferindo na Inteligência Artificial. Acho que a IA pode ajudar em muita coisa, mas eu acho que a questão é saber selecionar os dados e saber se aqueles dados são confiáveis né. Acho que isso é o mais importante. Não a IA em si, mas quem selecionou os dados para ser utilizado na IA.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por que?

Sim, eu acho. Por exemplo, a oncologia é muito usada para tratamento baseada em estudos que você tem os dados resposta, a partir do momento que você tiver mais dados, dados genéticos do paciente, dados clínico, mais a IA vai poder cruzar isso e dar, já faz né, quais são as melhores opções de tratamento de quimioterapia. Então eu acho que vai mudar muito na seleção do tratamento do paciente, enfim, eu acho que vai ajudar muito a ter um diagnóstico mais certeiro né. Mas não acho que, como eu estava falando, vai continuar tendo atendimento médico, mas acho que a inteligência artificial vai dar mais

segurança para os médicos.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico?

Em relação ao método de explicação? Eu acho que eles são, eles são excelentes. Assim, em relação a talvez o que vá Talvez eu não sei como é que a gente pode fazer isso...Esse aí é em relação ao diagnóstico, mas como é que esses modelos vão se encaixar na seleção do tratamento? Por exemplo, se você tiver um paciente oncológico, a gente seleciona um tipo de quimioterapia para saber na verdade, ou então se aquele paciente vai responder? Qual a chance de responder e de repente ele dá uma previsão de sobrevida e de risco de recorrência de doença. Não só o dado de ter câncer ou não. Como seria essa análise, com fatores, por exemplo, você ver risco de recidiva? Sobrevida do paciente? Toxidade? Se daria em todas as idades? Enfim é isso.

A. Entrevista PO12

Data: 12/02/2020

Onde: Hospital OncoDor / Barra da Tijuca

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método LIME?

Eu acho que...pelo o que eu entendo, o que ele consegue atribuir atribuir um valor maior ou menor para alguns fatores de risco né, e na presença ou ausência deles né, somar tudo isso e determinar o que a gente faz um pouco de cabeça né. Que é tentar entender o contexto dos exames e entender se a pessoa está com um risco maior ou menor de ter o câncer. E então como proceder em investigações posteriores ou qualquer coisa nesse sentido. Eu não saberia informar, qual desses fatores de risco tem um peso maior ou menor ai a presença ou ausência deles aumentaria ou diminuiria a chance de câncer. Meio que pra fazer as somas dos fatores no total. Ehhh, mas, mais, mas eu acho que é interessante por isso assim. Da gente tentar entender e quantificar um pouco melhor o risco individual de cada pessoa baseado nas coisas que nós normalmente avaliamos mesmo e tentar entender a questão da urgência de alguém ter a necessidade de fazer exames ou não.

Eu não sei por exemplo se já existe isso definido, por exemplo. Qual o nível de probabilidade de câncer que te faz buscar o exame X ou Y? Ou fazer ou não? Investigações posteriores ou alguma coisa nesse sentido, acho que são coisas que vale a pena avaliar e assim tentar validar né. porque pegar um paciente com 91% é um risco muito alto né, mas se pegar um paciente com um risco de 70%? Vale a pena fazer exames mais invasivos? Vale a pena acompanhar? São coisas que a gente às vezes ver em outras patologias usando alguns outros fatores. Acho que é um pouco isso.

Estamos testando a clareza da explicação dada pelo modelo acima. De acordo

com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Foi o Teste de Schiler, a colposcopia e o Exame de Papanicolau né? Acho que foi isso né?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo LIME é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Acho que sim, ehhh. Eu acho que sim, porque é um pouco da forma que a gente avalia uma questão de risco ou não. Eu só não sei se o termo correto é fez o diagnóstico né. Pelo que eu entendi ele dá a probabilidade daquela soma ser o diagnóstico. Mas eu acho que entendi bem sim. Por exemplo o teste de Schiller positivo a chance de ser é muito maior né? O Teste de Schiller positivo é como tá aqui no gráfico né? O tamanho, tamanho da barrinha né, a força é muito maior. Se a pessoa fez Colposcopia a chance de ele ter um câncer ou qualquer coisa é muito maior né. Então me pareceu claro sim.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo LIME? Que informação seria interessante ser apresentada neste método?

Assim, coisas a se acrescentar. Acho que, eu não sei se ele usa isso como outras formas de coisas a se ver. Por exemplo, tem outras DSTs, mas HIV é um fator de risco bastante importante para câncer de colo de útero, sei se ele leva em conta isso aí. Mas não sei, mas de uma forma geral me pareceu claro. A gente na nossa área da oncologia, uma coisa que a gente se preocupa muito de entender o risco de a pessoa ter câncer ou não, o que a gente faz nos nossos estudos clínicos é tentar, através dos dados, pesar que tipo de fator de risco é mais preponderante ou não é. Mas eu ficaria assim como é feito o balanço da importância de cada fator desse aí para somar e no final da conta dar esse risco. Mas me pareceu bastante razoável.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método SHAP?

É a mesma paciente? Deu 91 % né? Foi o mesmo do outro né? Em rosa os fatores que influenciaram né, ignorância minha, mas parece talvez por meio diferente mas a ideia é a mesma. Se atribuir o risco a cada fator desse né. E com isso pesar e definir melhor qual o risco de câncer que essa pessoa tem. Eu acho que os fatores são mais ou menos os mesmos, graficamente para o oncologista é muito mais fácil entender o teste de LIME,

porque é um gráfico muito parecido com gráfico de PlosPlot que a gente ver muito quando vai fazer os estudos clínicos , pra gente é mais fácil bater o olho e entender o que está querendo dizer. Mas eu acho que é isso assim. Uma coisa que eu vejo aqui além da questão gráfica, é porque lá no outro lá (LIME) ele me mostra aquele quadro lá, um pouquinho quanto pesa cada fator. Aqui, pelo menos pra mim não está tão claro. Aquele a gente consegue entender um pouco melhor. Mas e o valor está o mesmo né, isso quer dizer que existe uma semelhança. Mas não sei, eu não vejo tanta diferença assim, talvez por ignorância mesmo.

Estamos testando a clareza da explicação dada pelo SHAP. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Teste de Schiler, a colposcopia e o Exame de Papanicolau.

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo SHAP é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Sim. Isso ficou bastante claro...porque ele cita em parte o que a gente usa né, o Exame de Papanicolau. A única coisa que acho difícil e diferente é que enquanto algumas coisas que para a gente conta o resultado qualitativo e eu acho que ele (SHAP) leva em conta ter feito ou não. Por exemplo Papanicolau e Teste de Schiller. Porque existe uma série de coisas que a gente avalia no Exame de Papanicolau, uma série de outros resultados, a gente leva em conta, até pra confiar se o exame foi bem feito ou não né. E avaliar os tipos de alteração que tem né, o Teste Schiller também. Acho que que existe uma série de questões qualitativas que o método não consegue pegar com tanta facilidade. Perante a nossa cabeça e raciocínio clinico, isso também leva em conta a questão de risco maior ou menor. Mas acho que talvez transformar isso difícil, mas que de uma forma geral ficou bem claro o que é levado em conta para chegar nesse resultado de 91%.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo SHAP? Que informação seria interessante ser apresentada neste método?

Não sei se eu entendi errado, mas no LIME tem aquele quadro do lado que é o valor que dá, o peso que ele dá para cada fator de risco. E isso aqui na verdade é o número de parceiros que ela teve né? La no outro método (LIME), naquele quadro embaixo, ele dá o peso que ele leva em conta. Acho que de fato se isso proceder né, essa minha ideia. Acho que ter essa informação pra gente é importante, até pra gente entender o peso de cada fator

desse para chegar a esse resultado de 90% de risco e a gente até tentar comparar isso com o nosso raciocínio clínico, se esse resultado procede com o que a gente está pensando. Acho que talvez esteja sentindo um pouco de falta disso. Mas não sei se seria tão fácil fazer isso.

Baseado no que você viu acima. Como você explicaria o resultado provido pelo sistema "Dr.Inteligência Artificial"pelo método Permutation Importance?

Eu não entendi muito bem... Ah tá entendi, Mas não ficou tão claro assim para mim porque nos outros dois métodos você vê uma questão gráfica da porcentagem né e esse aqui é um pouco mais difícil de ler. Mas de uma forma geral, em ordem decrescente né? Do maior para o menor né? Eu me lembro de cabeça ali né, mas coloca o Teste de Schiller como o mais importante para chegar a esse resultado. E não sei né, mas talvez a ordem esteja diferente dos outros né? Mas agora eu estou entendendo melhor, e eu acho que a ideia é a mesma né. Mas talvez a apresentação dos outros métodos sejam mais fácil de entender, mas eu acho que a ideia é pouco a mesma.

Estamos testando a clareza da explicação dada pelo Permutation Importance. De acordo com esse modelo, quais as três características que mais influenciaram o diagnóstico de câncer de colo de útero da paciente?

Acho qie foi o Teste de Schiller, Idade da 1º relação e números de gestações. Não é isso?

O quanto você concorda com a seguinte afirmação: "A explicação gerada pelo modelo Permutation Importance é compreensível e me ajudou a entender como o sistema Dr.Inteligência Artificial diagnosticou o câncer da paciente.".

Sim, agora olhando melhor. Eu entendi melhor esse gráfico. Sim, foi bastante claro. No nosso raciocínio clínico é difícil a gente concordar assim. Se de fato isso tem um risco maior ou menor, o primeiro ou segundo, mas no geral o método é bem elucidativo sim.

Estamos testando a capacidade explicativa dos modelos. Que elementos você sentiu falta na explicação gerada por este modelo Permutation Importance? Que informação seria interessante ser apresentada neste método?

Acho que seria aquilo que eu falei no SHAP né. Mostrar melhor como chegou a esse resultado, porque cada fator desse está nessa ordem. Não fica muito claro, mesmo com o peso, como chegou nesse resultado. Qual a soma das informações que fazem chegar nesse valor aqui. Isso aqui já pé pre estabelecido né? Eu acho que nesse método, é que apresenta uma espécie de defeito, seria talvez a forma de visualização. A forma gráfica é

mais interessante pra gente entender. Aquela coisa de você bater o olho e já ter uma noção boa da informação. Talvez seja isso, esse aqui é necessário olhar com cuidado. Não tem informações percentuais como os outros, que dá pra gente entender bem melhor, o risco ou não. Quanto mais simples pra gente, mais fácil.

Qual método de explicação você mais gostou? Por quê?

O LIME em primeiro. Porque acho que pela informação gráfica ser mais clara. A gente, na pratica médica nós fazemos alguns tipos de exames que levam em conta muitas variáveis, então na apresentação da informação, existe sempre aquela questão dessa preocupação gráfica, pra facilitar essa leitura. Ao meu ver pelo menos, a forma de se apresentar os dados pelo LIME é mais fácil de ler. Uma coisa que você bate o olho e consegue entender melhor e mais rapidamente. Isso pra gente é muito importante.

Você confia na inteligência artificial/Aprendizagem de máquina? Por quê?

A princípio SIM. Tem áreas que isso está mais ou menos inserido, o nível de pesquisa e tudo. Mas de uma forma geral sim né, eu acho que na medicina o futuro vai envolver muita coisa relacionada a inteligência artificial, à machine learning. De uma forma geral nós confiamos. Eu acho que talvez só precisa fazer coisas, como você está fazendo né, que são estudos de como implementar os métodos. Mas eu tendo né, de antemão, sem discussões mais profundas a confiar em IA no geral.

Você acredita que a IA pode mudar e/ou revolucionar o domínio médico? Por quê?

Acho que sim. Sem dúvida. Porque medicina é lidar com dados né. Seja nos estudos ou na prática clínica diária. Quando o paciente senta na sua frente, o raciocínio clinico nada mais é do que você pegar diversas informações, dados que o paciente te da, seja contando histórias, seja resultado de exames e a gente juntar isso. A capacidade de raciocínio clinico é sua capacidade de juntar isso e entender o contexto e saber o que está acontecendo, o que não está. Pesar risco e benefício das coisas, fazer avaliação de risco. E acho que na minha impressão de leigo é que muito disso a IA consegue fazer.

Também é uma ignorância minha, mas eu acho que muito na interpretação disso, especialmente no contato com o paciente, ainda existe uma necessidade de determinar um grau de sensibilidade, não só emocional, mas da técnica mesma assim, que eu não sei se a IA consegue fazer. Mas é claro que nessa questão de ajudar a gerar dados, otimizar informações e coisa desse tipo, eu não tenho dúvida nenhuma que vai ser de importância fundamental. Já existem estudos em oncologia pelo menos, da utilização de IA, inclusive

para fazer diagnóstico de exame de imagem. Porque muita coisa de exame de imagem né, são reconhecimento de padrões. E existe muita variabilidade interpessoal nisso. Algumas pessoas são melhores em alguns métodos, outras pessoas em outros. Isso depende de uma habilidade individual e acho que muita dessa coisa de reconhecimento de padrões e tudo a IA consegue fazer bem né. Então sem dúvida existe um grande espaço. Existe uma discussão sobre isso né, até quando a IA consegue o humano e o médico, mas eu acho que isso nunca vai ser completo, mas o papel da IA eu tenho certeza que será cada vez maior. Pra tudo né, para pesquisa médica, para dados.

Que sugestão ou crítica você teria para que esses métodos sejam implantados no domínio médico ?

Sugestão ou crítica? Eu acho difícil assim né. Mas oncologista especialmente, é um ser, normalmente muito crítico e muito desconfiado com dados né. A nossa especialidade envolve muito isso né, você interpretar dados e talvez um esclarecimento maior quanto a como essas informações são avaliadas e pesadas para chegar a um número final. A forma como são feitos vai fazer com que uma pessoa vá confiar mais ou menos nos métodos e consequentemente aceitar mais e em última análise ser mais difundido. Crítica? Crítica é difícil assim. Mas é interessante assim né, tem vários empregos possíveis de se utilizar, mas eu acho que eu ficaria nessa situação mesmo. De tentar entender como se faz o peso de cada variável dessa para tentar como chegam nessa porcentagem, esses números.

Referências

[1] Data Science Academy, “Data Science Academy,” 2020.

[2] Data Science Academy, “Deep Learning Book,” 2020.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” 2016.

[4] S. Ramprasaath R., M. Cogswell, A. Das, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” 2017.

[5] F. Hohman, H. Park, C. Robinson, D. Horng, and P. Chau, “SUMMIT : Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations,” 2019.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-Agnostic Interpretability of Machine Learning,” no. Whi, 2016.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

[8] S. Russel and P. Norvig, *Artificial Intelligence A Modern Approach*. Pearson Education, third edit ed., 2016.

[9] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[10] O. Biran and C. Cotton, “Explanation and justification in machine learning: A survey.,” in *In IJCAI-17 Workshop on Explainable AI (XAI)*, pp. 8–13, 2017.

[11] D. Pedreschi, D. Informatica, U. Pisa, and L. B. Pontecorvo, "Discrimination-aware Data Mining," pp. 560–568, 2008.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *Leibniz International Proceedings in Informatics, LIPIcs*, vol. 67, pp. 1–23, 2017.

[13] M. Söllner, A. Hoffmann, H. Hoffmann, A. Wacker, and J. M. Leimeister, "UNDERSTANDING THE FORMATION OF TRUST IN IT ARTIFACTS," vol. 127, no. June, pp. 1–18, 2012.

[14] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," *FAT* 2019Guidotti - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 279–288, 2019.

[15] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," vol. 277, no. 2003, pp. 1–21, 2016.

[16] A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence," pp. 71–94, 2017.

[17] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pp. 80–89, 2019.

[18] C. Molnar, *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. 2019.

[19] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," no. Ml, pp. 1–13, 2017.

[20] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, pp. 50–57, Oct. 2017.

[21] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Broadway Books., 1 ed., 2016.

[22] D. Gunning, D. Io, and A. I. System, "Explainable Artificial Intelligence (XAI) The Need for Explainable AI," Tech. Rep. November, 2017.

[23] B. Herman, "The Promise and Peril of Human Evaluation for Model Interpretability arXiv : 1711 . 07414v1 [cs . AI] 20 Nov 2017," no. Nips, 2017.

[24] E. Thelisson, K. Padh, and L. E. Celis, "Towards Trust, Transparency, and Liability in AI/AS Systems," in *In IJCAI-17 Workshop on Explainable AI (XAI)*, pp. 53–57, 2017.

[25] S. Lundberg and S.-I. Lee, "An unexpected unity among methods for interpreting model predictions," no. Nips, pp. 1–6, 2016.

[26] S. M. Lundberg and S.-i. Lee, "A Unified Approach to Interpreting Model Predictions," in *Conference on Neural Information Processing Systems (NIPS 2017)*, no. Section 2, (Long Beach, CA, USA), pp. 1–10, 2017.

[27] S. M. Lundberg, G. G. Erion, and S.-i. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," no. 2, 2019.

[28] T. Parr, C. Turgutlu, Kerem Christopher, and J. Howard, "Beware Default Random Forest Importances," 2018.

[29] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, pp. 1–11, 2008.

[30] K.-F. Lee, *Inteligência artificial*. Rio de Janeiro: Globo Livros, 1 ed., 2019.

[31] T. M. Mitchell, *Machine Learning*. USA: McGraw-Hill, Inc., 1 ed., 1997.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[33] T. Mitchel, V. Podgorelec, and M. Zorman, "Decision Tree Learning," in *Encyclopedia of Complexity and Systems Science*, ch. 3, pp. 1–28, 2015.

[34] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A Survey Of Methods For Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, pp. 1–45, 2018.

[35] Z. C. Lipton, "The Mythos of Model Interpretability," no. Whi, 2016.

[36] A. A. Freitas, "Comprehensible Classification Models – a position paper," vol. 15, no. 1, pp. 1–10, 2014.

[37] T. Wang, E. Jones, F. Doshi-Velez, Y. Liu, E. Klampf, and P. MacNeille, "A Bayesian Framework for Learning Rule Sets for Interpretable Classification," *Journal of Machine Learning Research*, vol. 18, pp. 1–37, 2017.

[38] B. Letham, C. Rudin, T. H. Mccormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," vol. 9, no. 3, pp. 1350–1371, 2015.

[39] H. Yang, C. Rudin, and M. Seltzer, "Scalable Bayesian Rule Lists," pp. 1–31, 2017.

[40] M. Alber, S. Lapuschkin, P. Seegerer, and M. Hagele, "iNNvestigate neural networks!," 2018.

[41] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-dickstein, "SVCCA : Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability," no. Nips, pp. 1–17, 2017.

[42] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," 2017.

[43] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network Dissection : Quantifying Interpretability of Deep Visual Representations," in *Computer Vision and Pattern Recognition*, 2017.

[44] Y. Zharov and A. Tuzhilin, "YASENN: Explaining Neural Networks via Partitioning Activation Sequences," 2018.

[45] M. Robnik-Šikonja and M. Bohanec, "Chapter 9 Perturbation-Based Explanations of," in *Computational Economics Human and Machine Learning*, pp. 159–175, Springer International Publishing, 2018.

[46] D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods," no. Whi, 2018.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pp. 1527–1535, 2018.

[49] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, pp. 1189–1232, 10 2001.

[50] D. W. Apley, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," pp. 1–36, 2016.

[51] A. Goldstein, A. Kapelner, and J. Bleich, “Peeking Inside the Black Box : Visualizing Statistical Learning with Plots of Individual Conditional Expectation,” pp. 1–22, 2014.

[52] M. Staniak and P. Biecek, “Explanations of model predictions with live and breakDown packages,” vol. XX, pp. 1–16, 2018.

[53] A. Fisher, C. Rudin, and F. Dominici, “Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the "Rashomon"Perspective,” 2018.

[54] P. Biecek, “DALEX : Explainers for Complex Predictive Models in R,” *Journal of Machine Learning Research 19*, vol. 19, pp. 1–5, 2018.

[55] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, “Learning to Explain : An Information-Theoretic Perspective on Model Interpretation,” in *International Conference on Machine Learning*, 2018.

[56] B. Kim and U. T. Austin, “Examples are not Enough , Learn to Criticize ! Criticism for Interpretability,” in *Conference on Neural Information Processing Systems (NIPS 2016)*, no. Nips, 2016.

[57] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[58] P. W. Koh and P. Liang, “Understanding Black-box Predictions via Influence Functions,” in *International Conference on Machine Learning*, 2017.

[59] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, “Xai—explainable artificial intelligence,” *Science Robotics*, vol. 4, no. 37, 2019.

[60] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, “Trends and Trajectories for Explainable, Accountable and Intelligible Systems,” pp. 1–18, 2018.

[61] D. Doran, S. Schulz, and T. R. Besold, “What does explainable AI really mean? A new conceptualization of perspectives,” *CEUR Workshop Proceedings*, vol. 2071, 2018.

[62] S. Mohseni, N. Zarei, and E. D. Ragan, “A survey of evaluation methods and measures for interpretable machine learning,” *arXiv preprint arXiv:1811.11839*, 2018.

[63] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, pp. 210–215, 2018.

[64] I. Lage, E. Chen, J. He, M. Narayanan, and B. Kim, "An Evaluation of the Human-Interpretability of Explanation arXiv : 1902 . 00006v2 [cs . LG] 28 Aug 2019," pp. 1–24, 2019.

[65] A. Páez, "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)," *Minds and Machines*, vol. 29, no. 3, pp. 441–459, 2019.

[66] S. Mohseni and E. D. Ragan, "A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning," 2018.

[67] H. J. P. Weerts and W. V. Ipenburg, "A Human-Grounded Evaluation of SHAP for Alert Processing," 2019.

[68] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, and U. States, "Designing Theory-Driven User-Centric Explainable AI," in *Proceedings of the 2019 CHI conference on human factors in computing systems.*, (Glasgow, Scotland, UK), pp. 1–15, 2019.

[69] J. M. Corbin and A. Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc, 3rd ed., 2007.

[70] A. M. Nicolaci-da Costa, "O campo da pesquisa qualitativa e o Método de Explicitação do Discurso Subjacente (MEDS)," *Psicologia: Reflexao e Critica*, vol. 20, no. 1, pp. 65–73, 2007.

[71] Instituto Oncoguia, "O que é oncologia?," 2020.

[72] I. N. D. CÂNCER, "Câncer do colo do útero," 2020.

[73] Kelwin Fernandes, Jaime S. Cardoso and J. Fernandes., "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," in *Iberian Conference on Pattern Recognition and Image Analysis*, Springer International Publishing, 2017.

[74] A. M. Nicolaci-da Costa, D. Romão-Dias, and F. Di Luccio, "Uso de entevistas online no método de explicitação do discurso subjacente (MEDS)," *Psicologia: Reflexao e Critica*, vol. 22, no. 1, pp. 36–43, 2009.

[75] A. M. Nicolci-da Costa, C. F. Leitão, and D. Romão-Dias, "Como conhecer usua´rios atrave´s do Me´todo de Explicitaça˜o do Discurso Subjacente (MEDS)," *VI*

Simpósio sobre Fatores Humanos em Sistemas Computacionais - Mediando e Transformando o Cotidiano., pp. 47–56, 2004.