Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

Jefferson da Costa Lima

Desafios para a adoção de Inteligência Artificial pelo Sistema Único de Saúde (SUS): ética, transparência e interpretabilidade

Rio de Janeiro
2022

Jefferson da Costa Lima

Desafios para a adoção de Inteligência Artificial pelo Sistema Único de Saúde (SUS): ética, transparência e interpretabilidade

Tese apresentada ao Programa de Pós-Graduação em Informação e Comunicação em Saúde do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) para obtenção do grau de Doutor em Ciências.

Orientador: Dr. Marcel de Moraes Pedroso
Co-orientador: Dr. Christovam Barcellos

Rio de Janeiro
2022

Lima, Jefferson da Costa.

Desafios para a Adoção de Inteligência Artificial pelo Sistema Único de Saúde (SUS): ética, transparência e interpretabilidade / Jefferson da Costa Lima. - Rio de Janeiro, 2022.
146 f.

Tese (Doutorado) – Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Pós-Graduação em Informação e Comunicação em Saúde, 2022.

Orientador: Marcel de Moraes Pedroso.
Co-orientador: Christovam Barcellos.

Bibliografia: f. 109-118

1. Ciência de Dados. 2. Saúde Pública. 3. Inteligência Artificial. 4. Aprendizagem de Máquina. 5. Ética e Interpretabilidade. I. Título.

Elaborado pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/Icict/Fiocruz com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Igor Falce Dias de Lima - CRB-7/6930.

Fundação Oswaldo Cruz
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Programa de Pós-Graduação em Informação e Comunicação em Saúde

Jefferson da Costa Lima

Desafios para a adoção de Inteligência Artificial pelo Sistema Único de Saúde (SUS): ética, transparência e interpretabilidade

Aprovado em 19 de julho de 2022.

Dr. Marcel de Moraes Pedroso (Orientador)
(LIS/PPGICS/ICICT/Fiocruz)

Dr. Christovam Barcellos (Co-Orientador)
(LIS/PPGICS/ICICT/Fiocruz)

Dr. Ricardo Antunes Dantas de Oliveira
(LIS/PPGICS/ICICT/Fiocruz)

Dr. Rodrigo Murtinho de Martinez Torres
(LACES/PPGICS/ICICT/Fiocruz)

Dr. Flavio du Pin Calmon
(John A. Paulson SEAS/Harvard)

Dr. Alexandre Dias Porto Chiavegatto Filho
(LABDAPS/FSP/USP)

Dr. Fábio André Machado Porto
(DEXL/LNCC)

Rio de Janeiro
2022

Agradecimentos

Aos meu filhos, Fernanda e Gabriel, e à minha esposa, Katia, pelo incentivo e apoio nas inúmeras decisões que me trouxeram até aqui.

À minha mãe, irmãos, sobrinhos, tias e tios, que de tantas formas diferentes contribuíram para a minha formação.

Aos meus orientadores, Marcel Pedroso e Christovam Barcellos, pela generosidade na condução desse processo, pela dedicação e pela confiança depositada.

Às contribuições dos membros da banca de avaliação para o aprimoramento deste trabalho. Obrigado Rodrigo Murtinho (PPGICS/ICICT/FIOCRUZ), Ricardo Dantas (PPGICS/ICICT/FIOCRUZ), Flávio du Pin Calmon (SEAS/HARVAD), Alexandre Chiavegatto Filho (LABDAPS/FSP/USP) e Fábio Porto (DEXL/LNCC).

Aos inúmeros professores que, ao longo do tempo, me ajudaram a construir um olhar que tenta ser interdisciplinar e com o qual eu procuro analisar os desafios.

À equipe da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS), projeto do Laboratório de Informação em Saúde (LIS/ICICT) da Fiocruz, pela parceria nos esforços para compreender as melhores formas para aplicar Ciência de Dados no campo da saúde.

Aos amigos que muito me ajudaram nessa caminhada, especialmente aos membros da PCDaS Marcel Pedroso, Igor Morais e Lucas Carraro, por estarem sempre disponíveis e interessados nas discussões sobre o uso ético da inteligência artificial na saúde.

À melhor turma que eu poderia ter durante o doutorado, fonte de inspiração e inúmeras novas amizades.

Resumo

Nos últimos anos, aplicações que utilizam componentes baseados em Inteligência Artificial (IA) têm se tornado onipresentes em nosso cotidiano. No campo da saúde, ao mesmo tempo em que a IA pode promover enormes avanços, uma questão que se apresenta, e que é desafiadora para a sua adoção, é fazer com que o seu uso seja justo e não discriminatório contra pessoas, grupos, comunidades, populações e instituições. Outra característica marcante é que parte do êxito recente das aplicações baseadas em IA está relacionada a modelos cada vez mais complexos, sacrificando o entendimento humano sobre o seu funcionamento. Em áreas potencialmente sensíveis como a saúde, a falta de transparência é uma limitação que pode ocultar tratamentos discriminatórios ou, por falta de confiança na solução, funcionar como uma barreira para a adoção da tecnologia, o que pode levar a perda de enormes oportunidades para a melhoria do acesso aos serviços de saúde. Esta pesquisa, no primeiro momento, se concentra em identificar os fatores, técnicos, regulatórios e éticos, que podem contribuir para a construção de um ambiente mais adequado para o desenvolvimento da IA nos serviços de saúde. Em seguida, busca compreender como o funcionamento dos métodos de interpretabilidade, que podem fornecer, para modelos de IA reconhecidamente pouco transparentes (*black boxes*), alguma interpretabilidade e como isso pode impactar os serviços de saúde. Para isso, foi criado um experimento contrafactual relacionado a multiplicidade preditiva e consistência de explicações por meio de listas de importância dos atributos de quatro dos principais modelos de predição (classificação). Esta é uma pesquisa construída com um olhar para o Sistema Único de Saúde (SUS) e, especialmente, o seu princípio doutrinário da Equidade. Com isso, esperamos que o uso de IA na saúde possa contribuir para mitigar os efeitos das iniquidades sociais e econômicas, não para perpetuá-las ou agravá-las, ao criar um ambiente mais transparente e confiável.

Palavras-chave: Ciência de Dados. Saúde Pública. Inteligência Artificial. Aprendizagem de Máquina. Ética e Interpretabilidade.

Abstract

In recent years, applications that use Artificial Intelligence (AI) components have become ubiquitous in our daily lives. In the field of health, at the same time, when AI can make tremendous advances, a question that presents itself and is challenging for its adoption is to make its use fair and non-discriminatory against persons, groups, communities, populations, and institutions. Another striking feature is that part of the recent success of AI-based applications is related to increasingly complex models, sacrificing human understanding of their operation. In potentially sensitive areas such as health, lack of transparency is a limitation that can hide discriminatory treatments or, due to lack of confidence in the solution, act as a barrier to the adoption of technology, which can lead to the loss of enormous opportunities for improving access to health services, for example. This research first focuses on identifying the technical, regulatory, and ethical factors that can contribute to the construction of a more suitable environment for the development of AI in health services. Then, it seeks to understand how interpretability methods work, which can provide some interpretability for admittedly little transparent AI models (black boxes) and how this can impact health services. For this, a counterfactual experiment related to predictive multiplicity and consistency of explanations was created through lists of feature importances of four of the main prediction models (classification). This research was built with a look at the Unified Health System from Brazil (Sistema Único de Saúde - SUS) and its doctrinal principle of Equity. With this, we hope that using AI in health can mitigate the effects of social and economic inequities, not perpetuating or exacerbating them, by creating a more transparent and credible environment.

Keywords: Data Science. Public Health. Artificial Intelligence. Machine Learning. Ethics and Interpretability.

Lista de ilustrações

Lista de quadros

Lista de abreviaturas e siglas

AUC-ROC	*Area Under the Curve ROC*
COMPAS	*Correctional Offender Management Profiling for Alternative Sanctions*
DL	*Deep Learning*
EUA	Estados Unidos da América
EBM	*Explainable Boosting Machine*
Fiocruz	Fundação Oswaldo Cruz
GDPR	*General Data Protection Regulation*
HCBS	*Home & Community Based Services*
IA	Inteligência Artificial
IBM	*International Business Machines Corporation*
LAI	Lei de Acesso à Informação
LGPD	Lei Geral de Proteção de Dados
LIAA	*Laboratorio de Inteligencia Artificial Aplicad*
LR	*Logistic Regression*
MCTI	Ministério da Ciência, Tecnologia e Inovações
ML	*Machine Learning*
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
OMS	Organização Mundial da Saúde
RD	Retinopatia Diabética
SMOTE	*Synthetic Minority Oversampling Technique*
SUS	Sistema Único de Saúde
UTI	Unidade de Tratamento Intensivo
XAI	*Explainable Artificial Intelligence*

Sumário

1 Introdução

Nos últimos anos, o uso de *Machine Learning* (ML) e Inteligência Artificial (IA) têm assumido um papel de destaque em diversos campos. Seja na indústria, na economia ou na saúde, essas aplicações tornaram-se possíveis graças ao acúmulo de grandes volumes de dados sobre os mais variados aspectos da humanidade, dos indivíduos, do ambiente em que vivemos e das interações que nele ocorrem. Obviamente, processar esses dados, descobrir padrões ou aprender algo novo a partir deles só é possível em função da existência de capacidade computacional e de técnicas para o seu processamento.

Ao mesmo tempo, tudo leva a crer que o volume de dados continuará crescendo de forma acelerada e este cenário garante um ambiente bastante propício para a aplicação de *Machine Learning* em campos ainda mais diversos.

Na saúde, dada a sua relevância, há uma grande expectativa de que o uso de ML possa conduzir o setor à redução de custos, à ampliação do acesso e a uma maior precisão em diagnósticos e tratamentos. Além disso, o volume de investimentos globais no campo da saúde torna este um mercado relevante para a atuação de governos e empresas privadas, levando a preocupações regulatórias.

Esse esforço regulatório objetiva proteger a privacidade, a titularidade dos dados[1] e garantir transparência nas decisões algorítmicas. E, ao mesmo tempo, tentando não se tornar uma barreira para o desenvolvimento de uma tecnologia com potencial para trazer enormes ganhos à toda a sociedade. Obviamente, este é um equilíbrio complexo, ainda mais quando a necessidade de transparência no uso e tratamento de dados depara-se, dentre outras, com questões como a proteção de propriedade intelectual, as limitações técnicas para garantir interpretabilidade, a discussão sobre privacidade e as questões éticas envolvidas, com consequências que podem levar a um tratamento discriminatório contra pessoas, grupos, comunidades, populações e instituições.

Um ponto fundamental nessa discussão é a necessidade de termos algum nível de interpretabilidade sobre as decisões resultantes de modelos de *Machine Learning* (ML). Estes modelos têm melhorado continuamente seu desempenho preditivo, eventualmente se equiparando, ou até mesmo superando seres humanos para as mesmas atividades. Infelizmente, em grande parte, este processo tem sido baseado em algoritmos que geram modelos pouco transparentes (*black boxes*).

Nesse cenário, ganha impulso uma área de pesquisa que busca dar transparência aos modelos de ML opacos (*black boxes*) por meio da geração de artefatos compreensíveis por humanos. Esta área, em inglês, é conhecida como *Explainable Artificial Intelligence*

[1] Lei Geral de Proteção de Dados Pessoais (LGPD), Capítulo III - DOS DIREITOS DO TITULAR

(XAI). Em português, são usados termos como explicabilidade e interpretabilidade, que serão discutidos no capítulo 4.

Para modelos pouco transparentes, algum nível de interpretabilidade é mandatório em áreas fortemente reguladas, como o setor financeiro, pois não basta saber qual foi a predição do modelo, é importante que entendamos os motivos que levaram a ela e, quase sempre, quais mudanças seriam necessárias para alterar um desfecho desfavorável. Por exemplo, um empréstimo negado.

Para o campo da saúde, apesar das enormes promessas de aplicações de *Machine Learning*, é fundamental entender o que motivou uma determinada decisão. Não parece razoável que médicos e pacientes recebam sem questionar diagnósticos que não podem ser explicados. Neste caso, a preocupação e busca por soluções tecnológicas "interpretáveis" são fundamentais não só para dar transparência para médicos e pacientes, mas para todas as partes interessadas, inclusive aos órgãos reguladores.

Nesse sentido, dada a opacidade de alguns tipos de modelos de ML, a interpretabilidade assume um papel fundamental para muitos campos distintos. Para os desenvolvedores, torna-se possível entender o funcionamento interno de partes do modelo e, com isso, buscar melhorar o seu desempenho. Na pesquisa científica, muitas aplicações utilizam modelos baseados em redes neurais profundas (*Deep Learning*), que são por natureza pouco transparentes. Neste caso, a XAI pode desempenhar um papel importante ao dar acesso aos padrões identificados durante o processo de treinamento do modelo, o que pode ser útil na definição de novas hipóteses. Além desses atores, os mais diversos usuários do modelo podem ter interesses em explicações, das mais simples às mais complexas, o que destaca o fato de que a efetividade de uma explicação está relacionada à adequação ao público a que se destina.

> Primeiro, os detalhes ou as razões usadas para explicar dependem completamente do público ao qual são apresentados. Segundo, se a explicação deixou o conceito claro ou fácil de entender também depende completamente do público. Portanto, a definição deve ser reformulada para refletir explicitamente a dependência da explicabilidade do modelo em relação ao público[2] (Barredo Arrieta et al., 2020, p. 4).

Por outro lado, além de pessoas interessadas em entender como um modelo toma determinada decisão, a interpretabilidade assume também um papel fundamental na atribuição de responsabilidades por essas decisões (*accountability*) e na avaliação de sua conformidade a padrões e normas legais, técnicas e éticas. Já são muito bem documentados exemplos de aplicações de ML que apresentavam diversos vieses com relação a gênero,

[2] *First, the details or the reasons used to explain are completely dependent on the audience to which they are presented. Second, whether the explanation has left the concept clear or easy to understand also depends completely on the audience. Therefore, the definition must be rephrased to explicitly reflect the dependence of the explainability of the model on the audience.*

raça, idade, etc., tornando-se um elemento de perpetuação ou, até mesmo, ampliação de um tratamento discriminatório contra alguns grupos.

Em geral, o tratamento discriminatório por modelos baseados em ML tem como origem os dados utilizados em seu treinamento. Estes dados, caso capturem vieses culturais existentes na sociedade (racismo, machismo, idadismo, etc.), podem gerar soluções que reproduzam o tratamento discriminatório praticado nesta sociedade e, com isso, trazer enormes prejuízos a indivíduos ou grupos, funcionando como uma barreira no acesso a serviços e direitos e pode ser crítico em áreas como a saúde.

Neste ambiente, quando as soluções desenvolvidas são baseadas em modelos opacos de *Machine Learning*, a identificação e correção de tratamentos inadequados pode se tornar inviável, o que justifica o crescente interesse pela XAI.

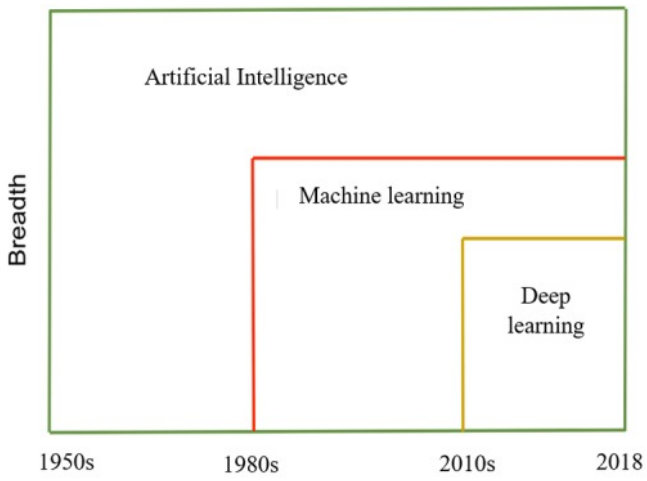
No entanto, a *Explainable Artificial Intelligence* (XAI), apesar de estar em forte desenvolvimento e ter papel central para criação de um ambiente eticamente mais seguro, sem tratamentos discriminatórios e confiável para aplicações de ML, ainda é uma área de pesquisa recente e em consolidação. Por isso, muito autores sugerem que as decisões baseadas em Inteligência Artificial, quando aplicadas a áreas potencialmente sensíveis, priorizem algoritmos que gerem modelos intrinsecamente explicáveis.

Por um lado, o principal argumento para os defensores do uso de modelos do tipo *black box* é o de que eles são mais precisos que outras abordagens, o que compensaria o custo de uma menor transparência. Entretanto, essa não é uma questão consensual entre os pesquisadores no tema. Por exemplo, Rudin e Radin (2019, p. 3) afirmam que é imprecisa a ideia de que se deva sacrificar a precisão quando se busca interpretabilidade. Segundo as autoras, essa ideia tem permitido que empresas vendam modelos complexos ou *black box*, quando há alternativas de modelos interpretáveis, mais simples para as mesmas tarefas.

Nesta tese, são discutidas questões éticas, técnicas e regulatórias que formam o ambiente em que estão sendo desenvolvidas algumas das principais soluções tecnológicas baseadas em *Machine Learning* (ML) no campo da saúde. Essa discussão visa a construção de um panorama que possa contribuir para uma compreensão mais ampla dos elementos envolvidos no campo em estudo, fomentando a criação de um ambiente mais justo e confiável para aplicações de ML.

Para tornar mais clara uma parte da terminologia utilizada nesta tese, a figura 1 apresenta a relação entre Inteligência Artificial, aprendizado de máquina (*Machine Learning*) e aprendizado profundo (*Deep Learning*), com a linha do tempo representada no eixo horizontal. Nesta tese, os três termos serão usados respeitando o exposto na figura, ou seja, Inteligência Artificial contém *Machine Learning* e este contém *Deep Learning*. Sempre que possível será utilizado o termo mais específico.

Figura 1 – Relação entre inteligência articicial, *Machine Learning* e *Deep Learning*



Fonte: Tobore et al. (2019)

1.1 DEFINIÇÃO DO PROBLEMA

Esta tese busca compreender alguns dos requisitos necessários para construir um ambiente mais seguro para o uso de IA, com foco no Sistema Único de Saúde (SUS), sob o olhar do seu princípio doutrinário da Equidade. Esta pesquisa busca caminhos que possam aumentar a probabilidade de que a IA na saúde contribua para mitigar os efeitos das iniquidades sociais e econômicas, não para perpetuá-las, criando um ambiente mais transparente para os setores vulneráveis da nossa sociedade.

É importante destacar que o SUS está imerso em uma sociedade com desigualdades históricas e graves, por isso, é lamentável, mas também razoável, esperar que alguns dos seus efeitos se manifestem em sua estrutura e construção histórica, por exemplo, em uma distribuição desigual de recursos, insumos e serviços.

Como um exemplo, em um estudo realizado por Boccolini et al. (2016), com dados da Pesquisa Nacional de Saúde de 2013, é feita uma avaliação dos fatores associados à discriminação percebida nos serviços de saúde do Brasil. Segundo o autor, 10,5% da população brasileira relatou sentir-se discriminada nos serviços de saúde, “sendo a falta de dinheiro (5,7%) e classe social (5,6%) as mais frequentemente apontadas”.

Com isso em mente, a preocupação principal desta pesquisa é a de contribuir para o desenvolvimento de um ambiente responsável para o uso da IA no SUS, que torne mais transparentes as decisões, ao invés de ocultar comportamentos possivelmente discriminatórios com aspectos relacionados à complexidade algorítmica.

1.2 OBJETIVOS DO ESTUDO

Objetivo geral

O objetivo desta pesquisa é o de analisar e discutir os fatores técnicos, éticos e regulatórios que possam contribuir para a criação de um ambiente responsável para o uso de IA no setor saúde, com foco no Sistema Único de Saúde (SUS), sob o olhar do seu princípio doutrinário da Equidade.

A primeira parte da pesquisa (capítulos 2 e 3) busca mapear oportunidades, riscos e limitações para um uso justo da IA. São analisados fatores técnicos, regulatórios e éticos, com um relato de caso e revisão da literatura.

Na segunda parte (capítulos 4 e 5), o foco se concentra em fatores técnicos que podem contribuir para aumentar a transparência dos componentes baseados em IA, para isso, emprega-se a revisão da literatura e a construção de um experimento contrafactual.

Objetivos específicos

1. Identificar os fatores técnicos, regulatórios e éticos que possam influenciar a criação do ambiente para uso de IA no SUS;
2. Analisar a literatura e discutir um caso real de aplicação de IA, com impacto na saúde, e relacioná-lo com a discussão desta pesquisa;
3. Analisar e comparar as propostas para garantir interpretabilidade em soluções tecnológicas que utilizam IA;
4. Realizar experimento contrafactual para explorar a consistência e estabilidade de métodos de interpretabilidade;
5. Analisar e discutir os possíveis impactos dos resultados do experimento com os métodos de interpretabilidade;
6. Definir e propor recomendações para a discussão sobre a construção de um ambiente justo e seguro para a adoção de IA no SUS;
7. Contribuir com o debate sobre o marco regulatório de IA em saúde no país.

1.3 JUSTIFICATIVAS

Relevância Teórica

Esta pesquisa caracteriza-se como um estudo que busca identificar oportunidades, riscos e limitações para o uso de IA no campo da saúde, contribuindo para a construção

de um ambiente justo e seguro para a aplicação da tecnologia no SUS.

Este estudo inova ao apresentar uma visão integrada, em que parte dos fatores que influenciam a solução tecnológica final, técnicos, regulatórios e éticos, são vistos como componentes de uma solução maior. Nesse sentido, a discussão pode contribuir para a compreensão dos limites da IA, quando vista isolada dos fatores externos, para tornar este ambiente justo e não discriminatório, pois esses fatores externos podem moldar profundamente o desenho das soluções tecnológicas.

Um outro aspecto importante das aplicações de IA é a opacidade de parte dos modelos utilizados atualmente, mesmo em algumas decisões sensíveis. Essa questão pode impactar enormemente as pessoas afetadas, pois essa opacidade torna difícil assegurar que uma determinada decisão foi justa e, no campo da saúde, a transparência é um princípio fundamental para garantir a segurança da aplicação. Por isso, esta pesquisa analisa e discute algumas das principais abordagens que se propõem a fornecer alguma interpretabilidade sobre as decisões algorítmicas, especialmente para aquelas tomadas por modelos do tipo *black box*.

Relevância Prática

Como fator motivador para esta pesquisa, as minhas atividades profissionais estão ligadas ao Sistema Único de Saúde (SUS), como servidor público federal em uma instituição de informação científica e tecnológica em saúde. Atuo como Tecnologista em Saúde Pública no Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz (Icict/Fiocruz), que tem como missão participar da formulação, implementação e avaliação de políticas públicas, desenvolver estratégias e executar ações de informação e comunicação no campo da ciência, tecnologia e inovação em saúde, objetivando atender às demandas sociais do Sistema Único de Saúde (SUS) e de outros órgãos governamentais.

Além disso, dada a sua importância econômica e social, a saúde é um campo relevante para a aplicação de *Machine Learning* (ML). Só nos Estados Unidos da América (EUA), as despesas nacionais com saúde chegaram a US$ 4,1 trilhões em 2020 (National Health Expenditure, 2021). Ao mesmo tempo, os sistemas de saúde têm enfrentado uma pressão por gastos crescentes, dentre outros motivos, pelo envelhecimento da população. As aplicações de IA/ML são uma aposta para a diminuição de custos nos serviços de saúde, mas a introdução da tecnologia traz consigo enormes desafios.

Um deles é a opacidade, frequentemente apontada nas soluções baseadas em IA. Ela pode estar ligada à complexidade envolvida, a preocupações com a privacidade dos dados ou, ainda, a demandas ligadas à proteção de propriedade intelectual. Entretanto, garantir um processo transparente e justo não se restringe às questões técnicas. Elementos

como a regulação, a ética e, até mesmo, a diversidade das equipes envolvidas, podem influenciar a tecnologia final.

Por outro lado, compreendendo possíveis impactos do uso de modelos pouco transparentes em decisões sensíveis, muitos esforços têm sido dirigidos para fornecer alguma interpretabilidade para os modelos opacos, geralmente com custos computacionais cada vez maiores e, além disso, com dificuldades para tornar as explicações confiáveis. Segundo Rudin (2019), no lugar de "tentar criar modelos que sejam intrinsecamente interpretáveis, houve uma explosão recente de trabalho sobre ML explicável".

A abordagem nesta tese tenta compreender como esses diversos elementos, técnicos, regulatórios e éticos, interagem para a criação de uma solução tecnológica baseada em IA. Em seguida, reconhecendo a importância da transparência para setores como o da saúde, esta pesquisa discute os principais métodos propostos para fornecer interpretabilidade e alguns riscos associados com o seu uso, quando não se opta por métodos intrinsecamente interpretáveis.

1.4 FORMULAÇÃO DO PROBLEMA

Em função dos objetivos elencados acima e da justificativa sobre sua relevância, podemos listar as principais questões de pesquisa a serem respondidas pela tese.

1.4.1 Questões de Pesquisa

1. Ao buscar a construção de um ambiente o mais confiável e justo possível para a aplicação de *Machine Learning* na Saúde, quais aspectos devem ser observados e quais atores devem estar presentes?

2. As estratégias selecionadas, propostas para a interpretabilidade de modelos de *Machine Learning* são adequadas para aplicações no campo da Saúde?

1.4.2 Hipóteses de pesquisa

A primeira hipótese a ser examinada pela pesquisa busca compreender os elementos que influenciam a construção de um ambiente, no contexto da saúde, que possa ser justo e responsável para o uso de IA. A hipótese apresentada é a de que a solução tecnológica baseada em IA pode ser decisivamente condicionada por fatores externos à tecnologia e, assim, garantir um uso ético pode depender mais desses fatores do que da própria IA.

- Hipótese 1: O desenvolvimento de um ambiente confiável e justo para a aplicação de *Machine Learning* na Saúde inclui as questões tecnológicas, mas não está restrito a elas.

A segunda hipótese discute a possibilidade de que métodos de interpretabilidade distintos possam gerar explicações substancialmente diferentes, mesmo quando aplicados a casos idênticos. Essa situação, pode minar a confiança nos modelos de IA e impedir a adoção da tecnologia, além de criar barreiras para que um paciente, por exemplo, possa identificar os reais motivos que levaram a um desfecho, ou, possivelmente, compreender que mudanças podem levar a um desfecho diferente.

- Hipótese 2: Métodos de interpretabilidade distintos podem gerar explicações substancialmente diferentes, mesmo quando aplicados a casos idênticos.

2 Relato sobre Estudo de Caso: Predição de gravidez precoce em Salta (Argentina)

A apresentação deste caso tem por objetivo discutir algumas das questões e preocupações envolvidas na adoção de soluções de Inteligência Artificial como parte de políticas públicas, especialmente as que se relacionam com o tema saúde. Os capítulos seguintes analisam em maior profundidade alguns aspectos e possíveis desafios que foram destacados neste estudo de caso.

Dentre outros, este caso foi selecionado em função da documentação disponível e da sua relação com o tema no Brasil e o possível impacto de projetos similares no SUS.

2.1 Apresentação

A adolescência corresponde ao período dos 10 aos 19 anos e é marcada por profundas mudanças, dentre as quais se destacam a conscientização da sexualidade, a estruturação da personalidade e a integração social (YAZLLE, 2006).

Segundo a Organização Mundial da Saúde (WHO, 2020), estima-se que 777 mil meninas entre 10 e 14 anos dão à luz a cada ano em países em desenvolvimento. Entre as que têm entre 15 a 19 anos, são 12 milhões de partos por ano. As complicações durante a gravidez e o parto são a principal causa de morte de meninas de 15-19 anos. Por outro lado, os bebês de mães adolescentes enfrentam maiores riscos de baixo peso ao nascer e parto prematuro, dentre outros problemas.

Além dos risco imediatos para a saúde da mãe e da criança, há efeitos de longo prazo que podem marcar profundamente o desenvolvimento psicossocial e econômico da mulher. A gravidez na adolescência, em muitos casos, leva ao abandono escolar, além de consequências sociais que podem incluir estigma, rejeição ou violência por parte de parceiros, pais e colegas (WHO, 2020). Segundo Yazlle (2006), a gravidez neste "grupo populacional vem sendo considerada, em alguns países, problema de saúde pública".

2.2 Inteligência artificial e prevenção de gravidez precoce na província de Salta (Argentina)

Em 2017, a província de Salta, na Argentina, firmou um acordo de colaboração com a Microsoft para usar Inteligência Artificial (IA) na prevenção da gravidez na adolescência e do abandono escolar (WHO, 2021; JEMIO; HAGERTY; ARANDA, 2022). Segundo News

Center Microsoft Latinoamérica (2018), os "algoritmos inteligentes permitem identificar características nas pessoas que podem levar a qualquer um desses problemas e alertar o governo para que ele possa trabalhar para preveni-los" (tradução nossa)[1]. O objetivo era identificar uma pessoa e associar a ela uma probabilidade de ocorrência de um desses desfechos, permitindo que intervenções preventivas pudessem ser aplicadas.

2.2.1 Coleta de dados

Para treinar o modelo de *Machine Learning* proposto, foram coletados dados junto à população de baixa renda, público alvo do programa. No total, 296.612 pessoas foram entrevistadas, das quais 12.692 eram mulheres entre 10 e 19 anos (Ortiz Freuler; IGLESIAS, 2018). Ao todo, a base era composta de 78 atributos (anexo A), que podem ser agregados seguintes grupos: pessoais, educação, saúde, trabalho, moradia e família. No quadro 1 são apresentadas algumas das questões levantadas.

Um ponto destacado por alguns analistas foi a falta de questões sobre o acesso a métodos contraceptivos ou o acesso à abordagem, no ambiente escolar, de temas ligados à educação sexual (TECNOPOLÍTICA, 2018). Pode-se também questionar se a lista de questões coletadas (anexo A), seria a mais adequada, pois focam quase exclusivamente em aspectos pessoais. Segundo Peña e Varon (2021):

> A abordagem da infância vulnerável é uma abordagem neoliberal clássica, aplicada por organizações como o Banco Mundial na região, e provém da ideia da pobreza como um problema individual (não sistêmico) e dos assistentes sociais como protetores de pessoas "em risco".

Outro item importante, que pode conduzir à subnotificação de casos, impactando a qualidade final do conjunto de dados reunido, é o estigma associado ao aborto e à gravidez precoce (LIAA, 2018), pois é razoável supor que a família ou a adolescente não informem gravidezes interrompidas, o que afetaria a qualidade do conjunto de dados. Dentre as pergunta feitas durante a coleta de dados, havia duas que tratavam de gravidez na adolescência da mãe ou de alguma irmã, além uma pergunta específica sobre a existência de gravidezes anteriores da pessoa entrevistada (ver anexo A)

2.2.2 Modelo desenvolvido

Segundo News Center Microsoft Latinoamérica (2018), o modelo desenvolvido tinha um nível de precisão de quase 90% (medido a partir de um teste piloto realizado em Salta). Pablo Abeleira, coordenador de Tecnologia do Ministério da Primeira Infância da Província de Salta, afirmou que:

[1] *"[...] algoritmos inteligentes permiten identificar características en las personas que podrían derivar en alguno de estos problemas y advierten al gobierno para que puedan trabajar en la prevención de los mismos."*

Quadro 1 – Dados coletados - predição de gravidez precoce

Grupo	**Atributo**
Pessoal	- Data de nascimento - Etnia - Estado civil - País de origem
Educação	- Nível educacional máximo alcançado - Há quanto tempo você está ausente de um estabelecimento de ensino?
Saúde	- Número de deficiências - Número de Doenças Crônicas - Teve gestações anteriores
Trabalho	- Trabalha - Trabalho não formal - Motivo de não trabalhar
Moradia	- Material do piso - Tem banheiro - Quantidade de banheiros
Família	- Idade da mãe - Tipo de trabalho da mãe - Estado civil da mãe - País de origem da mãe - Etnia da mãe - Nível educacional máximo alcançado pela mãe - Mãe engravidou na adolescência - Alguma irmã engravidou na adolescência

Fonte: Ortiz Freuler e Iglesias (2018, p. 35, Digital Annex, tradução nossa).

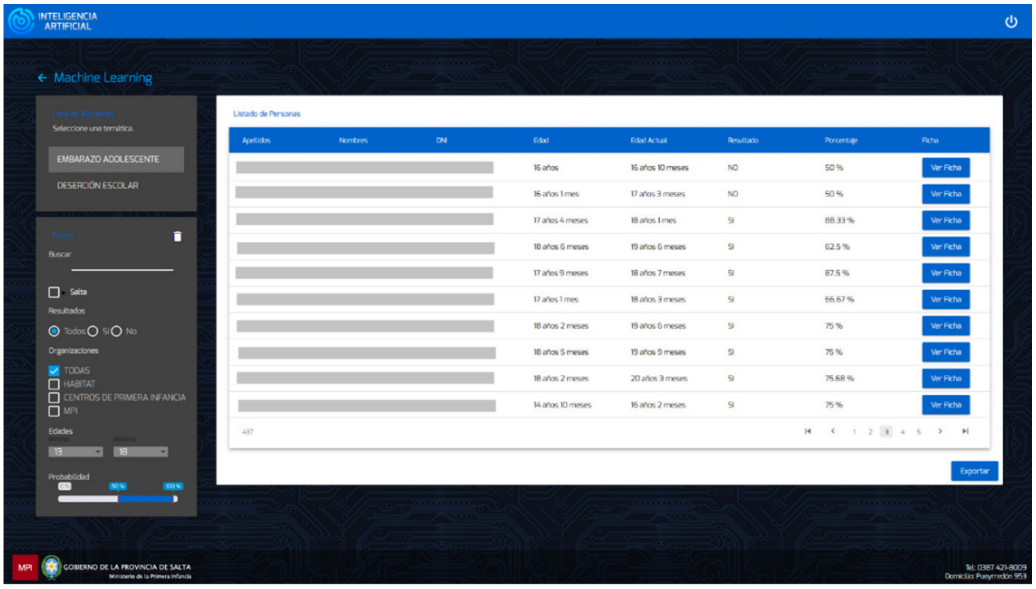
> [...] com esta informação, estamos nos organizando para fazer uma abordagem abrangente da situação com os recursos e capacidades que temos. Usamos soluções Microsoft Cloud para resolver situações sociais e dar às pessoas uma melhor qualidade de vida[2]. (News Center Microsoft Latinoamérica, 2018, tradução nossa)

Na figura 2 é possível ver uma imagem do painel de controle desenvolvido pelo Ministério da Primeira Infância de Salta. São exibidos os resultados para um grupo de mulheres e a probabilidade individual de ocorrer uma gravidez precoce.

A figura 3 exibe os dados de uma mulher identificada como tendo alta probabilidade de gravidez precoce. Segundo o governador da província de Salta em 2018, Juan Manuel

[2] [...] *con esta información, nos estamos organizando para hacer un abordaje integral de la situación con los recursos y las capacidades que tenemos. Usamos las soluciones en la Nube de Microsoft para poder resolver situaciones sociales y darles a las personas una mejor calidad de vida*

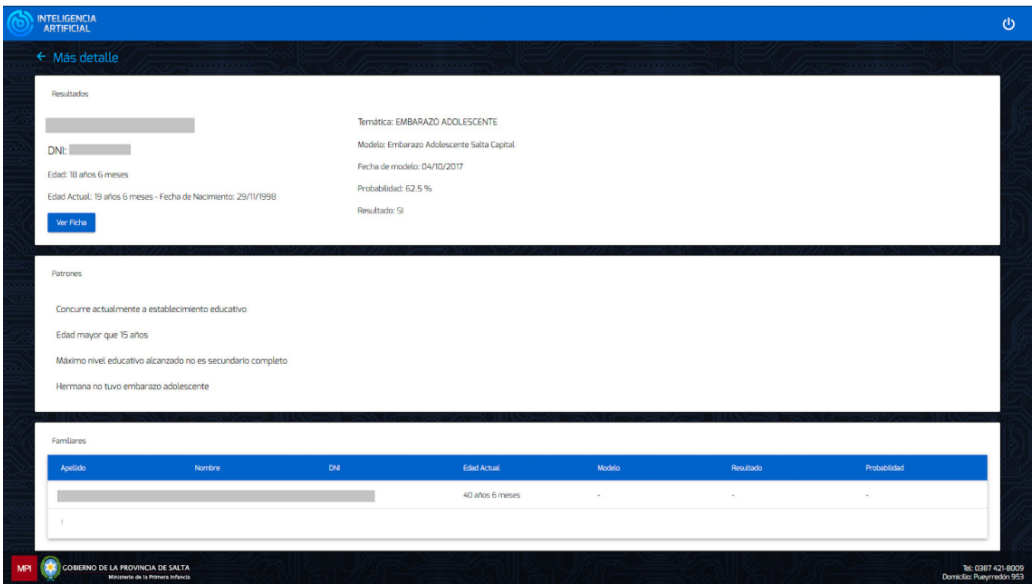
Figura 2 – Painel de controle - visão grupo



Fonte: Ortiz Freuler e Iglesias (2018)

Urtubey, "com a tecnologia é possível prever com cinco ou seis anos de antecedência, com nome, sobrenome e endereço, qual menina está 86% predestinada a ter uma gravidez na adolescência". [3](URTUBEY Y, 2018, tradução nossa)

Figura 3 – Painel de controle - visão indivíduo



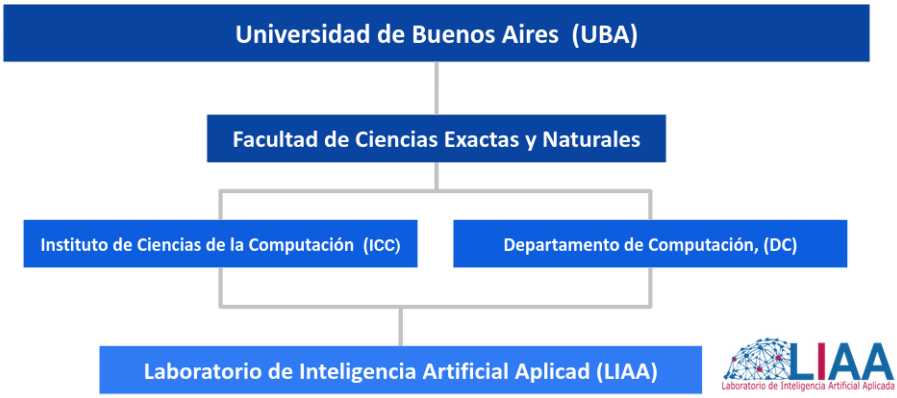
Fonte: Ortiz Freuler e Iglesias (2018)

[3] *"Con la tecnología vos podés prever cinco o seis años antes, con nombre, apellido y domicilio, cuál es la niña que está un 86% predestinada a tener un embarazo adolescente"*

2.3 Desafios identificados

No contexto da aplicação de Inteligência Artificial (IA) na prevenção da gravidez precoce em Salta (Argentina), várias análises apontam desafios, erros metodológicos e possíveis limitações na abordagem realizada (LIAA, 2018; Ortiz Freuler; IGLESIAS, 2018; WHO, 2020; PEÑA; VARON, 2021). Uma destas análises, feita em abril de 2018 pelo *Laboratorio de Inteligencia Artificial Aplicad* (LIAA) da *Universidad de Buenos Aires* (UBA) (ver organograma na figura 4). Nele são apontados três principais problemas: resultados superdimensionados, dados possivelmente tendenciosos e dados inadequados. Estes três itens são discutidos na seção 2.3.1.

Figura 4 – *Laboratorio de Inteligencia Artificial Aplicad* (LIAA) - organograma



Fonte: Elaborado pelo autor

2.3.1 Análise do *Laboratorio de Inteligencia Artificial Aplicad* (LIAA/UBA)

A análise feita pelo LIAA (2018) é dividida em três seções (problemas), que são abordadas a seguir, assim como as réplicas a cada uma delas enviadas pelo Ministério da Primeira Infância de Salta (Argentina), que podem ser encontradas em Ortiz Freuler e Iglesias (2018, p. 19).

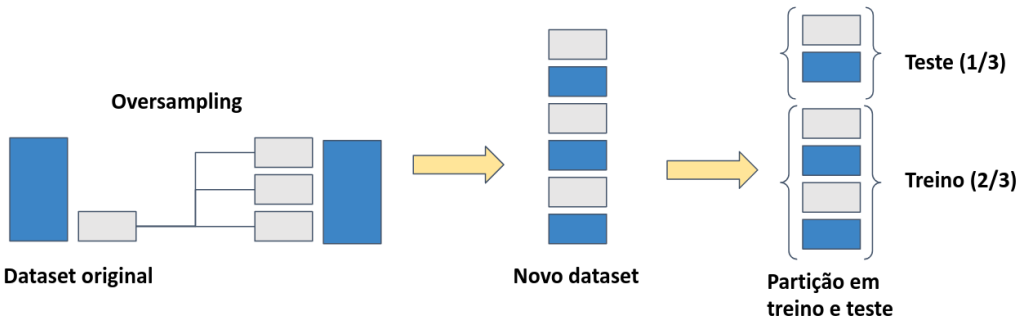
Problema 1: Os resultados superdimensionados estão relacionados a um erro metodológico, que foi possível identificar a partir de uma publicação no site GitHub.com feita por Facundo Davancens (DAVANCENS, 2017), funcionário da Microsoft na Argentina, detalhando alguns passos para a construção do modelo de predição de gravidez precoce. Infelizmente, este repositório não está mais disponível online (em 15/mar/2022), mas uma cópia do seu conteúdo, capturada durante a pesquisa para esta tese em 13/09/2020, pode ser consultada no anexo B.

Segundo a análise, o resultado do modelo estava provavelmente superdimensionado em decorrência de um problema conhecido como vazamento de dados (*data leakage*). Ele

surge antes do treinamento do modelo, durante a preparação dos dados, quando não é respeitado o princípio que requer que os testes de um modelo sejam feitos com dados que não tenham sido utilizados no seu treino.

Como apontado pelo LIAA, o processo de balanceamento, que foi realizado antes da separação dos dados que seriam utilizados para treino e teste do modelo, permite que os testes sejam feitos em dados já utilizados no treino (ver figura 5), o que superestima a precisão do modelo e impede que se possa fazer qualquer afirmação sobre a sua qualidade. Segundo LIAA (2018), a técnica utilizada para o balanceamento dos dados foi o *oversampling* com o método *Synthetic Minority Oversampling Technique* (SMOTE)[4], que gera novas amostras sintéticas baseadas na classe minoritária, mas com pequenas perturbações nos atributos originais.

Figura 5 – Problema 1: resultados artificialmente superdimensionados



Fonte: Elaborado pelo autor

Sobre este erro metodológico (Ortiz Freuler; IGLESIAS, 2018, p. 19), o Ministério da Primeira Infância de Salta (Argentina) confirma que o modelo publicado no GitHub reutilizou os dados de treinamento como dados de avaliação, mas que o modelo atual é baseado em testes usando dados independentes. O que, em resumo, indica que o problema teria sido corrigido.

Por outro lado, este ponto anuncia uma outra discussão que será abordada em inúmeros momentos ao longo deste trabalho: a importância da transparência nas aplicações de *Machine Learning*.

Problema 2: Os dados utilizados são, potencialmente, distorcidos, o que afeta a sua confiabilidade. Segundo o LIAA (2018), os dados sobre gravidez precoce podem ser tendenciosos ou incompletos, por se tratar de um assunto sensível e envolto em uma série de tabus. Em função disso, é razoável supor que alguns setores da sociedade, com mais acesso a recursos que podem incluir a interrupção da gravidez, acabem por distorcer os

[4] <https://docs.microsoft.com/pt-br/azure/machine-learning/component-reference/smote>

dados ao não informar essa interrupção. Consequentemente, mesmo que a metodologia utilizada para construir e avaliar os sistemas estivesse correta, os padrões identificados pelo modelo poderiam conduzir a conclusões equivocadas, refletindo distorções nos dados.

Sobre este problema (Ortiz Freuler; IGLESIAS, 2018, p. 19), o Ministério da Primeira Infância de Salta (Argentina) afirma não haver viés na base de dados porque o modelo busca fazer previsões apenas para essas populações vulneráveis, que seria o seu escopo de atuação.

Problema 3: Os dados coletados são inadequados para a tarefa de prever a gravidez precoce, pois, segundo LIAA (2018), os dados reunidos seriam capazes, na melhor das hipóteses, de determinar se uma adolescente teve ou tem uma gravidez. Afinal, é de se esperar que as condições e características de uma adolescente mude significativamente em cinco ou seis anos e, por isso, para prever a gravidez futura, seria importante mapear as mudanças no ambiente.

Logo, talvez seja insuficiente utilizar dados sobre moradia, trabalho, etc. de um ponto no tempo (ver anexo A). Como exemplo, o acesso facilitado a métodos contraceptivos e a introdução de disciplinas sobre orientação sexual no ambiente escolar podem ter um grande impacto nos índices de gravidez precoce, mas eles não eram monitorados nas variáveis informadas pelo Ministério da Primeira Infância.

Por outro lado, segundo Ortiz Freuler e Iglesias (2018, p. 19), o Ministério da Primeira Infância de Salta (Argentina) afirma que o "modelo é atualizado e retreinado regularmente com base em dados continuamente atualizados". Entretanto, esta resposta não deixa claro se serão coletadas sempre as mesmas informações, com que periodicidade, se será com o mesmo público e, em especial, se essas informações são eficazes para a predição da gravidez precoce.

Ortiz Freuler e Iglesias (2018, p. 19) destacam que seria importante a publicação de um "relatório técnico com informações sobre as bases de dados utilizadas e as variáveis selecionadas, bem como a hipótese que norteia o projeto do modelo e o processo que conduz ao modelo definitivo". Outro ponto importante destacado é a necessidade de comparações com outros modelos. Segundo os autores:

> Este documento explicativo também deve incluir uma análise de modelos semelhantes. Por exemplo, um estudo sobre abandono escolar precoce nos Estados Unidos descobriu que variáveis como etnia, educação especial, nascimento nos Estados Unidos, inglês como língua materna, comportamento escolar e região geográfica da escola não previam significativamente o abandono escolar precoce. No entanto, variáveis equivalentes foram incluídas no modelo de Salta. Espera-se que os responsáveis pelo design ofereçam uma explicação teórica para sua inclusão. [5] (Ortiz Freuler; IGLESIAS, 2018, tradução nossa)

[5] *"This explanatory document should also include an analysis of similar models. For example, a study on early school leaving in the United States found that variables such as ethnicity, special education,*

Certamente, para mitigar os riscos de impactos negativos na adoção de soluções como essa, a transparência assume um papel fundamental, mas ela deve vir acompanhada da formação de quadros técnicos que possam avaliar e monitorar os riscos envolvidos.

2.3.2 Outras análises: para além dos desafios técnicos

Dentre os textos utilizados na pesquisa para o desenvolvimento desta tese, os pontos destacados na análise feita pelo LIAA aparecem em quase todas as obras que abordam o mesmo caso. Entretanto, há outras questões relevantes, destacadas por outras autoras e autores, que podem ajudar a compreender melhor o contexto em que a solução tecnológica proposta é aplicada e que, de forma nem sempre clara, pode influenciar todo o desenho do projeto. Um exemplo marcante é a ausência dos adolescentes na discussão até este momento, com toda a responsabilidade sendo depositada sobre o "comportamento" feminino ou suas condições de vida. Segundo Joana Varón (apud Valente, 2020)

> "Além dos métodos estatísticos serem malfeitos, a iniciativa tem presunções sexistas, racistas e classistas sobre determinado bairro ou segmento da população. O trabalho é focado em meninas, somente, presumindo que os garotos não precisam aprender sobre direitos sexuais e reprodutivos. Temos que tomar cuidado para que segmentos já segregados não sejam mais discriminados sob uma máscara de opções neutras da tecnologia".

Sobre o contexto em que a solução de Salta é implementada, é importante lembrar que ela surge em meio à discussão sobre a descriminalização do aborto e, segundo Varon e Peña (2021, p. 18), Juan Manuel Urtubey, governador de Salta em 2018, político antiaborto e conservador, apresentou a iniciativa tecnológica como solução mágica. Para Sternik (2018), um dos objetivos era evitar a aplicação da lei de Educação Sexual Integral[6] e, segundo o autor (abril/2018), "até alguns meses atrás, Salta também foi a única província argentina que ofereceu educação religiosa em escolas públicas e privadas e não Educação Sexual Integral, apesar do que indica a Lei implementada há onze anos".

Outro ponto importante é a questão do consentimento informado para determinados usos dos dados de populações vulneráveis. Ainda sobre o caso de Salta, a Organização Mundial da Saúde (OMS), em publicação sobre o tema ética e governança de IA (WHO, 2021, p. 69), afirma que:

> "O algoritmo de predição também era inadequado, pois fornecia previsões que eram sensíveis para adolescentes sem seu consentimento (ou de seus pais), prejudicando assim sua privacidade e autonomia. Como o

birth in the United States, English as a mother tongue, school behaviour, and the school's geographical region did not significantly predict early school leaving. However, equivalent variables were included in the Salta model. Those in charge of design should be expected to offer a theoretical explanation for their inclusion."

[6] <https://www.argentina.gob.ar/educacion/esi/normativa>

> algoritmo visava indivíduos especialmente vulneráveis, era improvável que eles tivessem a oportunidade de contestar o uso das intervenções, e poderia reforçar atitudes e políticas discriminatórias". [7]

Por outro lado, Pablo Abeleira, coordenador de tecnologia do Ministério da Primeira Infância de Salta (apud Elebi, 2020, p. 22) afirma que os dados dos adolescente com menos de dezoito anos só são coletados com a concordância dos seus responsáveis, para os quais é explicado como a informação será utilizada.

Independente de haver ou não uma assinatura que formaliza a cessão dos dados pelos responsáveis, talvez o ponto mais relevante seja entender o quanto uma população tão vulnerável compreende os possíveis usos desses dados. Discutir o tipo de consentimento mais adequado e a melhor forma de proteger a população está fora do escopo desta tese, mas certamente deve fazer parte de qualquer projeto responsável e que lide com dados tão sensíveis.

É possível ver que, muito além dos desafios técnicos que impactam na efetividade da predição dos modelos, os autores citados nesta seção destacam questões que nem sempre são discutidas ao avaliar a adequação de um projeto com o uso de *Machine Learning*. Em especial, a influência do contexto, de preconceitos e de objetivos políticos no desenho da solução.

Como exemplo, uma visão machista sobre o problema da gravidez precoce fez com que os meninos não tenham nenhum papel definido. Não se pensou em estimar a probabilidade deles se tornarem pais nos próximos anos. Toda a responsabilidade é atribuída às meninas. Em outro contexto, o desenho proposto para este problema poderia ser outro.

Este, e muitos outros exemplos desta seção mostram que um problema pode ser enfrentado de várias formas diferentes e que usar IA não garante que a abordagem seja necessariamente neutra, imparcial ou objetiva.

2.4 Aplicações fora de Salta (Argentina)

Apesar de todas as questões apontadas por diversas análises, este projeto não ficou restrito à província de Salta na Argentina. Iniciativas similares, ou derivadas dela, foram avaliadas ou implementadas em outras províncias da própria Argentina (La Rioja, Tierra del Fuego, Chaco e Tucumán) e também em países da América Latina como o Brasil e a Colômbia (VARON; PEÑA, 2021; WHO, 2021; Ortiz Freuler; IGLESIAS, 2018).

[7] *"The predictive algorithm was also inappropriate, as it provided predictions that were sensitive for adolescents without their (or their parents') consent, thereby undermining their privacy and autonomy. As the algorithm targeted individuals who were especially vulnerable, it was unlikely that they would have the opportunity to contest use of the interventions, and it could reinforce discriminatory attitudes and policies"*

No Brasil, este projeto chega a partir de um acordo de cooperação técnica entre o Ministério da Cidadania, a província de Salta (Argentina) e a Microsoft (ver anexo C). Segundo este acordo, a proposta seria desenvolver uma "prova de conceito para implementar ferramentas de Inteligência Artificial que subsidiem melhoria das ações do programa Criança Feliz". O acordo de cooperação firmado em 23/09/2019 encerrou-se após os seis meses estabelecidos para a sua duração.

A cidade de Campina Grande na Paraíba foi a escolhida para os testes iniciais no Brasil (GOMES, 2009; G1-PB, 2019). No lançamento do projeto, conhecido como Projeto Horus, o então ministro Osmar Terra afirmou que "[...] essa parceria vai nos ajudar a melhorar o processo, melhorar a qualidade, fazer um atendimento maior e melhor para a população". No entanto, para além dos discursos positivos e aparentemente bem intencionados, o uso de IA com uma população tão vulnerável, e em um tema tão sensível, deve se apoiar nas melhores práticas e evidências disponíveis para mitigar os riscos envolvidos.

Como visto no caso de Salta na Argentina (capítulo 2), em função dos questionamentos levantados, a transparência torna-se um requisito fundamental para elevar a segurança na adoção dessa tecnologia. No Brasil, pesquisadores que buscam compreender como estas ferramentas estão sendo utilizadas, com quais dados e que papel assumem nos processos de tomada de decisão, frequentemente precisam recorrer à Lei de Acesso à Informação (LAI)[8,9,10].

No final de 2020, a LAI foi o caminho utilizado por Peña e Varon (2021) para tentar compreender como se daria a implementação do Projeto Horus no Brasil, mas após a conclusão do acordo de cooperação (fevereiro de 2020), parece não haver registro sobre as margens de erro ou informações sobre o resultado da prova de conceito, embora o Ministério da Cidadania informe ter repassado para a Microsoft dados que estavam no Sistema Único de Assistência Social (SUAS), Cadastro Único e CADSUAS, do antigo Ministério de Desenvolvimento Social (agora Ministério da Cidadania).

Em resumo, apesar de toda polêmica sobre o Projeto Horus na Argentina, o Brasil forneceu dados sensíveis de uma população vulnerável e, aparentemente, não obteve nenhum retorno.

[8] Lei de Acesso à Informação (LAI): <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>

[9] Exemplo de pedido de acesso à informação 1: <http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=71003129432202071>

[10] Exemplo de pedido de acesso à informação 2: <http://www.consultaesic.cgu.gov.br/busca/_layouts/15/DetalhePedido/DetalhePedido.aspx?nup=71004002319201904>

2.5 Considerações sobre o estudo de caso relatado

A gravidez na adolescência, por seus efeitos de curto, médio e longo prazos, representa um importante desafio para qualquer sociedade. Neste sentido, como o uso de *Machine Learning* (ML) e Inteligência Artificial (IA) têm assumido um papel de destaque nos mais diversos campos, é razoável esperar por aplicações desta tecnologia no enfrentamento de eventos com impactos indesejáveis à saúde.

Entretanto, neste caso de Salta, é preciso entender que a tecnologia é utilizada em apenas um dos componentes de uma política pública mais ampla, que vai da concepção do projeto, passando pelos dados escolhidos para a coleta, até as ações estatais baseadas nas probabilidades associadas a cada adolescente. Logo, para afirmar que uma solução tecnológica ajuda a transformar uma política em algo mais justo ou efetivo, precisamos compreender as suas limitações, pois a tecnologia é incapaz de corrigir erros metodológicos ou dados enviesados, o que pode levar a tratamentos discriminatórios.

No contexto brasileiro, um tratamento discriminatório contra pessoas ou grupos vai de encontro ao princípio da equidade do Sistema Único de Saúde (SUS). A equidade tem relação direta com igualdade e justiça e busca reconhecer as diferenças nas condições de vida, saúde e nas necessidades das pessoas. No entanto, quando se fala em verificar se um tratamento é ou não justo, estamos discutindo uma questão que está ligada a inúmeros fatores, cuja comprovação sempre depende de um processo tão transparente quanto possível. Por outro lado, utilizar um processo pouco transparente, que não consegue demonstrar o quão adequado é, exige das pessoas afetadas um confiança cega em um processo opaco, o que é indesejável pelos riscos assumidos. Resta, a quem propõe a solução tecnológica, demonstrar que as vantagens superam os riscos.

Por isso, para que se possa aproveitar da melhor forma possível todo o potencial benéfico das mais diversas tecnologias relacionadas com o conceito de Inteligência Artificial, é importante a definição de um arcabouço regulatório específico. Este arcabouço deve assegurar a transparência adequada, além de buscar soluções que não tragam tratamentos discriminatórios ou que possam causar danos injustificáveis a pessoas ou grupos, principalmente os mais vulneráveis, que encontrariam dificuldade para reverter seus efeitos negativos.

Este estudo de caso aponta diversas questões que podem conduzir a conflitos com alguns dos princípios basilares do SUS, o que pode impactar negativamente seus usuários, criando desconfiança e impedindo que o sistema de saúde brasileiro possa fazer o melhor uso da IA. Na visão de Peña e Varon (2021):

> "[...] podemos dizer que a 'Plataforma Tecnológica de Intervención Social' e o Projeto Horus são apenas um exemplo bastante eloquente de como a pretensa neutralidade da Inteligência Artificial tem sido cada vez mais implementada em alguns países da América Latina para apoiar políticas

> públicas potencialmente discriminatórias que poderiam prejudicar os direitos humanos das pessoas sem privilégios, bem como para monitorar e censurar as mulheres e seus direitos sexuais e reprodutivos".

Por fim, como podemos ver em muitas áreas distintas, há inúmeros casos em que a Inteligência Artificial (IA) está sendo empregada com sucesso, mas como o Projeto Horus demonstra, usar Inteligência Artificial (IA) não torna o projeto necessariamente imparcial, justo ou eficaz, pois a sua efetividade depende de muitas varáveis externas que podem influenciar as saídas do modelo, gerando um tratamento discriminatório.

3 *Machine Learning* na saúde: oportunidades e desafios éticos, técnicos e regulatórios

Este capítulo se relaciona com a primeira pergunta de pesquisa desta tese. Com ela busca-se compreender o ambiente que cerca o desenvolvimento, uso e implantação dos sistemas baseados em IA e os elementos que podem tornar essas aplicações de *Machine Learning* mais confiáveis e justas, em especial no campo da saúde, observando os princípios do Sistema Único de Saúde (SUS).

3.1 Introdução

Nos últimos anos, aplicações que utilizam técnicas de *Machine Learning* (ML) têm se tornado onipresentes em nosso cotidiano. Elas estão disponíveis em celulares, computadores, meios de transporte, serviços de saúde e sistemas de segurança, por exemplo. Entretanto, apesar do seu uso em situações com grande impacto nas vidas das pessoas, nem sempre é possível compreender razoavelmente como determinadas decisões são tomadas por essas aplicações.

Por que tive o meu pedido de financiamento negado? Por que fui identificado como suspeito pela polícia? Por que tive acesso negado a um serviço de saúde? Estas são algumas explicações que as pessoas afetadas podem desejar, em especial quando a avaliação é feita por sistemas automatizados ou com apoio deles.

Em função da sua complexidade interna ou de restrições no acesso aos dados utilizados, modelos de ML costumam ser pouco transparentes e muitos sistemas são vendidos ou fornecidos como "caixas-pretas". Além disso, destacam-se também as restrições ligadas à propriedade intelectual ou as que se justificam pela proteção da privacidade. Modelos opacos ou sem transparência razoável tornam complexa a tarefa de desenvolver um ambiente confiável para as diversas partes interessadas, o que pode se tornar uma barreira para a sua adoção.

Além disso, aplicações de ML possuem características distintas ou mais acentuadas que outras tecnologias recentes. Algumas delas são a relativa autonomia para encontrar padrões a partir dos dados, a possibilidade de modificar o seu comportamento adaptando-se à medida que novos dados são fornecidos e, como já destacado anteriormente, uma certa opacidade oriunda da complexidade interna dos modelos. Essas características têm levado a uma crescente busca por técnicas que permitam alguma interpretabilidade sobre funcionamento interno dos modelos.

Para atender a essa demanda, nos últimos anos um grande esforço tem sido feito no desenvolvimento de um campo de estudos denominado *Explainable Artificial Intelligence* (XAI). Segundo Vilone e Longo (2021a, p. 1), a opacidade criou a necessidade de desenvolvimento das arquiteturas XAI em busca de novos métodos capazes de expor e explicar a lógica seguida por modelos de aprendizado de máquina. Explicabilidade, que é utilizada por alguns autores como sinônimo de interpretabilidade, é um tema central dentre os desafios técnicos para um uso seguro de ML e será abordado em maior profundidade no capítulo 4.

Quanto aos desafios regulatórios, em 2018 a União Europeia aprovou o Regulamento Geral de Proteção de Dados (GDPR). Nela, segundo Guidotti et al. (2018, p. 2), há um aspecto inovador sobre a tomada de decisão automatizada, que, em certa medida, garante o direito de obter "explicações significativas da lógica envolvida" quando ocorre a tomada de decisão totalmente automatizada, o que torna central o papel da XAI. Ainda segundo o autor, é necessário o desenvolvimento de tecnologias que consigam explicar a lógica que conduz a uma determinada decisão, pois, caso contrário, "corremos o risco de criar e usar sistemas de decisão que realmente não entendemos", e o impacto pode ser enorme.

Visando mitigar esses impactos, diversas iniciativas têm impulsionado o desenvolvimento de diretrizes, tecnologias e metodologias que permitam conferir alguma interpretabilidade sobre as decisões de modelos de ML. Grandes empresas, governos e a sociedade civil têm se envolvido neste esforço. Embora não haja consenso sobre muitos pontos propostos pelas diversas iniciativas, é possível ver que isso não tem impedido avanços importantes.

Ao mesmo tempo, além das questões técnicas e regulatórias, temos que entender o impacto que esses artefatos[1] tecnológicos podem representar do ponto de vista ético. Por exemplo, a empresa de consultoria Gartner previa que "até 2018, metade das violações da ética empresarial ocorrerão por meio do uso impróprio de análises de Big Data" (GUIDOTTI et al., 2018, p. 2).

Nesta tese, a seção 2.5 apresentou desafios éticos que surgiram no caso sobre a prevenção de gravidez precoce em Salta (Argentina) e, para alguns desses desafios, entender o funcionamento interno do modelo poderia ajudar na identificação de vieses. Na seção 3.3, serão apresentados outras situações que poderiam ter seus riscos mitigados caso houvesse uma maior compreensão dos limites das soluções propostas. Neste sentido, do ponto de vista técnico, espera-se que a XAI consiga cumprir o papel de dar mais transparência, permitindo enfrentar as questões éticas já colocadas e as que serão provocadas pela adoção de novas aplicações. Além disso, espera-se que a XAI desempenhe um papel fundamental

[1] Diversos autores vêm se dedicando ao desenvolvimento da temática sobre engenharia das decisões (no sentido simoniano de artefatos) buscando a produção de regras e estratégias que viabilizem sua operacionalização tanto do ponto de vista substantivo (se os resultados da decisão são bons ou positivos), quanto procedimental (se procedimentos que avaliam as consequências foram observados e se as escolhas alcançaram os resultados esperados (SIMON, 1970 apud Pedroso, 2011, p. 32)

no atendimento de requisitos regulatórios e no aprimoramento de sistemas baseados em IA.

Neste contexto, este projeto se propõe a analisar algumas questões éticas envolvidas, em especial as que podem ter impacto no campo da saúde e as iniciativas técnicas e regulatórias que têm sido desenvolvidas para o seu enfrentamento.

3.2 Oportunidades para a Inteligência Artificial na saúde

Em função da sua importância social e econômica, a saúde é um campo de destaque para a aplicação de *Machine Learning* (ML). Só nos Estados Unidos da América (EUA), as despesas nacionais com saúde chegaram a US$ 4,1 trilhões em 2020 e a projeção para 2028 é atingir o valor de US$ 6,2 trilhões (National Health Expenditure, 2021). Neste ambiente de gastos crescentes, há muitas apostas de que aplicações de ML possam ajudar a reduzir custos e melhorar a saúde da população.

Neste sentido, as promessas são de ampliação do acesso aos serviços de saúde, diagnósticos mais precisos, desenvolvimento de uma medicina personalizada e redução de tempo e custos para a descoberta de novos tratamentos, entre outras. Além disso, todo esse potencial se torna ainda mais importante em um momento de aumento da longevidade da população, com esperado impacto negativo nos custos dos sistemas de saúde.

Embora estejamos apenas no início do desenvolvimento de soluções de IA aplicadas ao campo da saúde, já é possível afirmar que algumas propostas são promissoras. Liu et al. (2019) realizou uma revisão sistemática em estudos que comparam o desempenho diagnóstico de modelos de aprendizagem profunda (*Deep Learning*) com o de profissionais da saúde ao analisar imagens médicas. A conclusão foi a de que o desempenho diagnóstico dos modelos de aprendizagem profunda é similar ou superior ao dos profissionais de saúde.

Muitas são as soluções que utilizam ML para apoiar a prática de profissionais de saúde (LIU et al., 2019, p. 2), como o exemplo anterior, mas já há exemplos de soluções autônomas que emitem o diagnóstico sem a necessidade de supervisão médica.

> [...] sistemas de IA autônomos em saúde são sistemas de IA que tomam decisões clínicas sem supervisão humana. Esses sistemas de IA de diagnóstico médico rigorosamente validados são uma grande promessa para melhorar o acesso aos cuidados, aumentar a precisão e reduzir custos, ao mesmo tempo que permitem que médicos especialistas forneçam o maior valor ao gerenciar e tratar pacientes cujos resultados podem ser melhorados[2]. (ABRÀMOFF; TOBEY; CHAR, 2020, p. 1, tradução nossa)

[2] [...] *autonomous AI systems in healthcare are AI systems that make clinical decisions without human oversight. Such rigorously validated medical diagnostic AI systems hold great promise for improving access to care, increasing accuracy, and lowering cost, while enabling specialist physicians to provide the greatest value by managing and treating patients whose outcomes can be improved.*

Seguindo esta abordagem, em 2018 foi aprovado o primeiro equipamento para diagnóstico de retinopatia diabética pela *US Food and Drug Administration* (FDA), o IDx-DR[3]. O objetivo era construir um dispositivo que pudesse ser usado "na atenção primária em ambientes sem requisitos específicos, por operadores sem experiência anterior em imagens de retina e com treinamento mínimo" (ABRÀMOFF; TOBEY; CHAR, 2020, p. 4).

Além das aplicações ligadas às análises de imagens, apoiando ou substituindo profissionais de saúde, há inúmeras oportunidades em atividades como o diagnóstico, prognóstico, tratamento e monitoramento de pacientes (QAYYUM et al., 2021, p. 4-7), no apoio à formação de profissionais de saúde (BEEDE et al., 2020, p. 7), na vigilância epidemiológica (BUDD et al., 2020, p. 1-3), no desenvolvimento de medicamentos (STEPHENSON et al., 2019) e no atendimento na atenção primária com o uso de assistentes virtuais (BAKER et al., 2020).

A seguir são apresentados alguns casos com o objetivo de destacar benefícios potenciais com aplicações de ML na saúde.

3.2.1 Detecção precoce de retinopatia diabética

A retinopatia diabética (RD) é um importante desafio para a saúde pública. Segundo Gonçalves Escarião et al. (2008), "as complicações microvasculares do diabetes mellitus na retina constituem-se na principal causa de cegueira da população economicamente ativa no Brasil e no mundo", ainda segundo o autor, "o tratamento dessas alterações é eficaz na prevenção da cegueira quando instituído precocemente".

Neste sentido, para contribuir com a detecção precoce da RD, muitos estudos têm sido desenvolvidos e algumas aplicações já se encontram disponíveis (ABRÀMOFF; TOBEY; CHAR, 2020; ELEBI, 2020; WHO, 2021).

Beede et al. (2020) descreve um estudo feito na Tailândia que utiliza aprendizagem profunda (*Deep Learning*) para a detecção de doenças oculares diabéticas. Como, nos estágios iniciais a RD é assintomática, o ministério da saúde do país estabeleceu uma meta anual de examinar 60% das pessoas com diabetes. Entretanto, com 4,5 milhões de pacientes e 1.500 oftalmologistas, dos quais apenas 200 são especialistas em retina, "a escassez de médicos limita a capacidade de triagem de pacientes e também cria um atraso de tratamento para aqueles que têm DR" (BEEDE et al., 2020, p. 1).

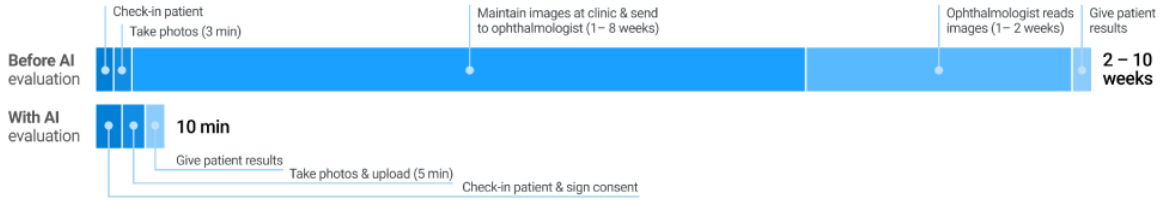
Para lidar com esse desafio, foi desenvolvido um algoritmo que faz a avaliação da retinopatia diabética, evitando a espera de semanas pelo resultado e pela revisão por um oftalmologista. O algoritmo demonstrou possuir uma precisão similar a de um especialista

[3] IDx-DR: <https://dxs.ai/products/idx-dr/idx-dr-overview/>

(superior a 90% de sensibilidade e especificidade), com uma redução de de 23% na taxa de falsos negativos e um leve aumento de 2% de falsos positivos (BEEDE et al., 2020, p. 2).

No processo tradicional, o exame para detecção de RD é realizado por um profissional de enfermagem, a imagem é armazenada e posteriormente enviada a um oftalmologista para análise. Até que o paciente tenha acesso ao resultado, este processo pode levar de duas a quatro semanas. Com o uso do algoritmo desenvolvido, este tempo pode ser reduzido para apenas dez minutos. A figura 6 ilustra a diferença entre as duas abordagens (com e sem a IA).

Figura 6 – Triagem ocular antes e depois da implantação do sistema de aprendizado profundo



Fonte: Beede et al. (2020)

Por fim, o uso da IA é uma grande esperança para superar as barreiras impostas pela falta de profissionais especializados e, nesse caso, fornecer resultados imediatos torna viável o início mais rápido do tratamento dos pacientes, o que é crucial no caso da retinopatia diabética.

3.2.2 Detecção de sepse

A sepse é um síndrome que exige um pronto reconhecimento e tratamento precoce. Ela pode ser definida com "uma síndrome de resposta inflamatória, causada por uma infecção que pode se originar em um local e causar alterações sistêmicas na tentativa de combatê-la" (RODRIGUES et al., 2022).

Segundo Kalil et al. (2018, p. 1, tradução nossa):

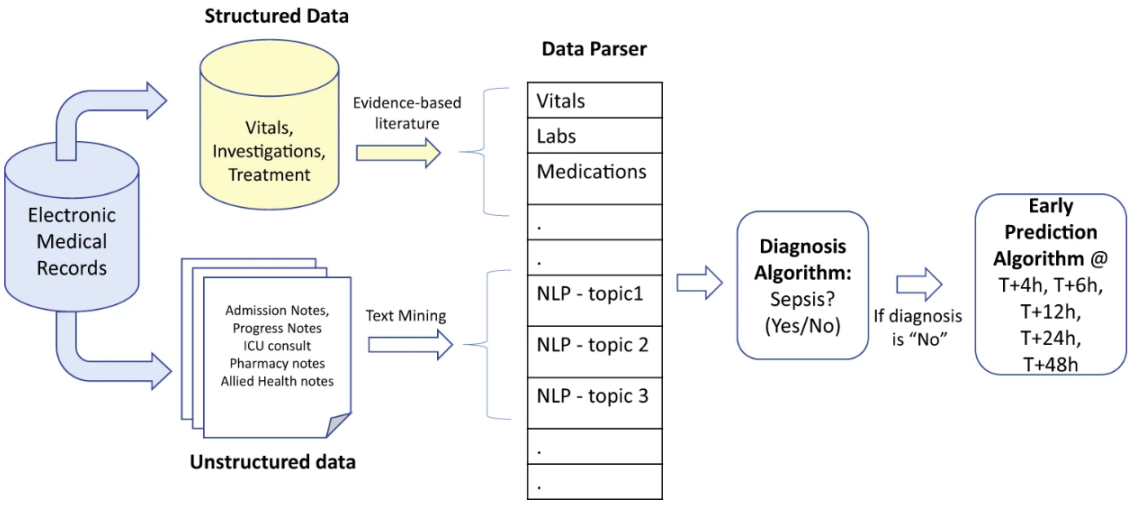
> [..] a sepse representa 25% das taxas de ocupação de leitos de UTI (Unidade de Terapia Intensiva) no Brasil e sua mortalidade associada pode variar de 29,6 a 54,1% em hospitais privados e públicos, respectivamente, tornando-se a doença de maior custo no setor saúde. O custo do atendimento de um paciente com sepse é seis vezes maior do que de um paciente sem sepse, com o custo aproximado de US$ 25.000 por paciente, totalizando US$ 17 milhões por ano.[4]

[4] [..] *sepsis accounts for 25% of ICU (Intensive Care Unit) bed occupancy rates in Brazil and its associated mortality can vary from 29.6 to 54.1z% in private and public hospitals, respectively, making it the costliest disease in the health sector. The cost of care for a patient with sepsis is six times higher than*

Os números deixam clara a importância da sepse nas taxas de mortalidade e o seu impacto financeiro. Com isso, muitos projetos têm avaliado soluções que incluem IA como um de seus componentes. Goh et al. (2021, p. 1) descrevem o projeto em que foi desenvolvido o algoritmo SERA, baseado em IA. Ele utiliza dados estruturados e notas clínicas não estruturadas para prever e diagnosticar sepse.

O objetivo do SERA é prever a ocorrência de sepse em pacientes com antecedência de 4, 6, 12, 24 e 48 horas, pois para essa síndrome o tratamento precoce é crucial para prevenir da mortalidade. A figura 7 mostra as etapas utilizadas para desenvolver o algoritmo SERA.

Figura 7 – Configuração do algoritmo SERA



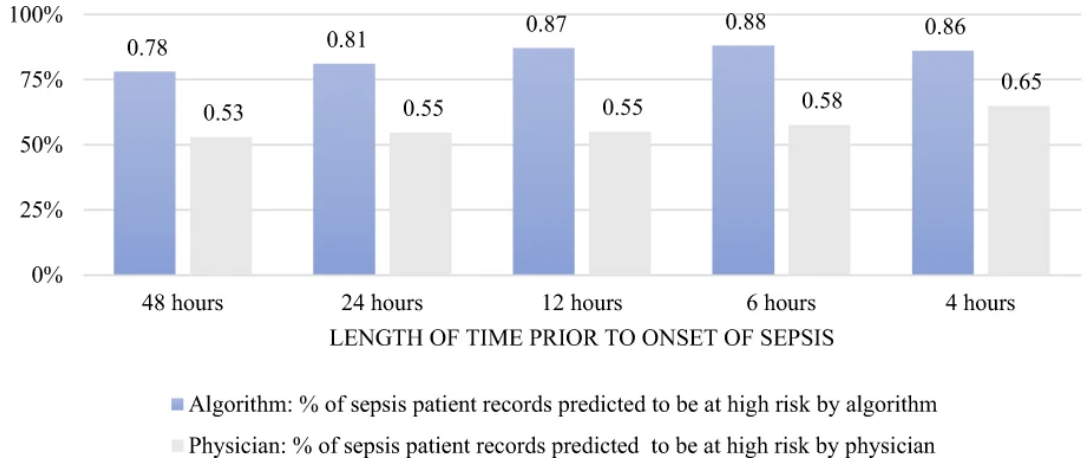
Fonte: Goh et al. (2021)

Quando comparado o desempenho do algoritmo com as previsões médicas para a incidência de sepse nos pacientes, foi verificado "um potencial do algoritmo para aumentar a detecção precoce de sepse em até 32% e o de reduzir os falsos positivos em até 17%" (GOH et al., 2021, p. 1). O desempenho do SERA alcançou uma precisão preditiva, 12 horas antes do início da sepse, para as métricas AUC (*Area Under the Curve*), sensibilidade e especificidade de 0,94, 0,87 e 0,87, respectivamente. A figura 8 compara a métrica sensibilidade entre o SERA e médicos.

Certamente, ainda há muito trabalho a ser feito para garantir a segurança e a efetividade das iniciativas que objetivam aplicar IA no campo da saúde. Entretanto, muitos trabalhos já apontam para resultados promissores, que continuarão a ser impulsionados nos próximos anos pelo desenvolvimento da Ciência de Dados, das infraestruturas tecnológicas e de um acesso crescente a dados com qualidade. Para que esses fatores convirjam em

of a patient without sepsis, with the approximate cost of $ 25,000 per patient, amounting to a total of $ 17 million per year.

Figura 8 – Comparação do desempenho entre o SERA e médicos para verdadeiros positivos (sensibilidade)



Fonte: Goh et al. (2021, p.5)

direção ao bem comum, será necessária a criação de um ambiente que seja adequadamente regulado e em que haja a devida proteção da privacidade e, ao mesmo tempo, seja o mais transparente possível.

Infelizmente, muitas aplicações de IA são suficientemente complexas para que nem os seus desenvolvedores sejam capazes de compreender a sua lógica interna. Com isso, essa falta de transparência pode se tornar uma barreira a sua adoção, pois o campo da saúde não pode prescindir da maior transparência possível sobre as decisões tomadas, ainda mais em situações potencialmente sensíveis.

As próximas seções deste capítulo discutem alguns dos desafios à adoção de soluções de IA na saúde. A questão da interpretabilidade, que busca dar maior transparência às decisões algorítmicas, será abordada em maior profundidade no capítulo 4.

3.3 Risco associados às decisões algorítmicas na saúde

Ao mesmo tempo em que *Machine Learning* (ML) na Saúde pode promover enormes avanços, uma questão que se apresenta, e que é desafiadora para a sua adoção, é fazer com que o seu uso seja justo e não discriminatório contra pessoas, grupos, comunidades, populações e instituições.

Como discutido no capítulo 2, sistemas de *Machine Learning* podem ser usados como um dos componentes de uma política pública, mas apesar de todo o rigor matemático ou estatístico que possa ser empregado, mesmo supondo que erros metodológicos não foram cometidos no tratamento dos dados, esses sistemas são incapazes de corrigir erros na concepção do projeto, dados de baixa qualidade ou de identificar e corrigir todas as

possibilidades de vieses nos dados. Com isso, pode ser gerado um efeito de ampliação ou perpetuação de tratamentos discriminatórios, questão agravada pela baixa transparências de alguns modelos.

Esta tese argumenta que a principal origem dos riscos pode ser externa ao modelo. Logo, é preciso um olhar mais amplo sobre o contexto em que o sistema é desenvolvido e, ao mesmo tempo, estar consciente das limitações típicas de soluções baseadas em Inteligência Artificial (IA). Com isso, busca-se uma visão mais equilibrada para a identificação das oportunidades em que a IA possa realmente contribuir para ampliar a eficiência, a eficácia e a efetividade dos sistemas de saúde, com foco na sua sustentabilidade e na qualidade dos serviços oferecidos, orientados pelos princípios de universalidade, integralidade e equidade do Sistema Único de Saúde (SUS).

3.3.1 Automação na alocação de benefícios para pessoas com deficiências

Nos Estados Unidos da América (EUA), o programa de saúde social Medcaid, por meio do serviço *Home & Community Based Services*[5] (HCBS), atende pessoas com deficiências intelectuais ou de desenvolvimento, deficiências físicas e/ou doenças mentais e oferece aos beneficiários a possibilidade de receber serviços de saúde em sua própria residência ou comunidade, em vez de instituições (hospitais ou asilos, por exemplo) ou outros ambientes que os distanciam da família e amigos.

Brown et al. (2020) e Lecher (2018) citam o caso de pacientes atendidos pelo HCBS no Arkansas (EUA). Eles têm paralisia cerebral e isso impacta significativamente sua mobilidade. Para tarefas como comer, vestir e mudar de posição, eles precisam contar com a assistência de outras pessoas. Dada essa condição, o HCBS cobria o custo de 56 horas semanais de cuidados para cada um deles. O número de horas era definido por uma enfermeira, com base na avaliação das necessidades do paciente, mas em 2016 este processo passou a ser orientado por um algoritmo que, pela descrição em Lecher (2018), parece ser uma árvore de decisão. Como consequência, sem nenhuma justificativa clara, o número de horas cobertos diminuiu de 56 para 32 horas semanais, embora as condições dos pacientes não tenham se alterado.

Neste caso, não é relevante discutir a tecnologia utilizada, ou até mesmo se há ou não o uso de IA nesta solução, pois o que se destaca é um padrão em que a decisão baseada na avaliação humana é trocada, sem a devida discussão, por um algoritmo que em tese seria mais objetivo. Lecher (2018) cita a justificativa dos gestores do serviço de que "[...] o sistema anterior, baseado em humanos, era propício ao favoritismo e a decisões arbitrárias"[6]. Entretanto, apesar das alegações de que as alterações no número de horas

[5] HCBS: <https://www.medicaid.gov/medicaid/home-community-based-services/index.html>

[6] *"[...] the previous, human-based system was ripe for favoritism and arbitrary decisions."*

financiadas derivam de uma maior objetividade do algoritmo, os usuários apontam que não há evidências que sustentem essas afirmações.

Em 2011, em um caso semelhante no estado de Idaho (EUA), uma solução construída internamente para alocar fundos de assistência domiciliar restringiu os benefícios de muitos usuários em até 42%. Mesmo assim, "[...] o estado se recusou a divulgar a fórmula que estava usando, dizendo que sua matemática se qualificava como segredo comercial"[7](LECHER, 2018). Em 2012, por meio de uma ação judicial que argumentava que esta situação limitava o direto dos usuários a recorrer de uma decisão, foi possível verificar que o algoritmo havia sido construído com dados inadequados, causando danos e, em função da falta de transparência, limitando enormemente a possibilidade de reversão.

Casos como o do Arkansas e de Idaho mostram a necessidade de alguma regulação que proteja a população em decisões sensíveis, em especial os grupos mais vulneráveis, dos efeitos negativos de soluções baseadas em algoritmos e, ao mesmo tempo, permita uma transparência adequada. Com o uso de *Machine Learning* essa questão atinge outro patamar. E é este ponto que impulsiona o desenvolvimento da *Explainable Artificial Intelligence* (XAI), tema que começa a ser discutido em maior profundidade na seção 3.4.

Em resumo, os casos de Arkansas e Idaho possuíam falhas técnicas, algumas identificadas e descritas em Brown et al. (2020) e Lecher (2018). Por outro lado, não havia uma regulação capaz de criar um ambiente adequado para a avaliação prévia desse algoritmo, o que, agravado pela falta de transparência, tornava complexo recorrer e até mesmo identificar erros sistêmicos.

3.3.2 Vieses e o papel dos conjuntos de dados para *Machine Learning*

Nesta seção serão discutidos casos que abordam riscos associados a alguns tipos de vieses nos conjuntos de dados, além da importância da adequada representação da população alvo nos dados de treinamento dos modelos de *Machine Learning* para mitigar riscos.

Sobre o espaço entre as oportunidades e os riscos associados ao uso de IA, Firth-Butterfield et al. (2022) afirma que:

> [...] embora a IA tenha muitos benefícios potenciais para nossa sociedade e o planeta, ela está longe de ser perfeita. Existem inúmeros casos de IA sendo usada, intencionalmente ou não, para excluir e enfraquecer indivíduos e comunidades, corroer os direitos humanos e minar nossas instituições democráticas[8].

[7] *"[...] the state declined to disclose the formula it was using, saying that its math qualified as a trade secret"*

[8] [...] *while AI holds many potential benefits for our society and the planet, it is far from perfect. There are numerous cases of AI being used, intentionally or unintentionally, to exclude and disempower individuals and communities, erode human rights, and undermine our democratic institutions.*

Mitigar os riscos passa por compreender a sua origem e pelo desenvolvimento de estratégias que possam reduzir seus efeitos prejudiciais. O objetivo é tornar os sistemas baseados em IA mais seguros e responsáveis ou, até mesmo, impedir o seu uso em situações para os quais os seus benefícios esperados não possam ser comprovados.

Muitas tarefas que podem ser desempenhadas por modelos baseados em Inteligência Artificial (IA) são possíveis graças aos dados fornecidos para o seu treinamento. O aprendizado supervisionado de *Machine Learning* busca a identificação de padrões em dados fornecidos como exemplos. Obviamente, a qualidade do modelo resultante está ligada à qualidade dos dados fornecidos, pois dados inadequados podem gerar modelos enviesados, que tratem pessoas de forma discriminatória, restringindo ou bloqueando o seu acesso a direitos e serviços.

Buolamwini e Gebru (2018) demonstram como algoritmos podem apresentar resultados bastante discrepantes com relação a raça e gênero. Na análise feita em 2017 pelas autoras, a taxa de erro na identificação do gênero a partir de fotografias era de 34,7% para mulheres negras e, no outro extremo, de apenas 0,8% para homens brancos. Foram avaliados três serviços de identificação de gênero (IBM, Microsoft e Face++), mas em todos eles os classificadores tinham melhor desempenho em rostos masculinos e peles mais claras. Este tipo de comportamento, quando essas ferramentas são integradas a soluções mais complexas, podem se tornar uma barreira contra grupos sub-representados nos dados ou que não são o foco dos processos de avaliação.

No campo da saúde, Morley et al. (2020) cita o exemplo dos monitores de frequência cardíaca que são menos precisos para aqueles com pele mais escura. Ruth Hailu (2019), descreve em seu texto os ajustes que podem ser feitos na intensidade ou no tipo de luz que deve ser utilizada para uma leitura mais precisa. Essa imprecisão torna-se preocupante quando lembramos que pessoas ou grupos são apenas dados para um modelo de *Machine Learning* e que os modelos gerados tendem a ter melhor qualidade com dados mais precisos.

O exemplo do monitor de frequência cardíaca, que podemos extrapolar por similaridade para outros equipamentos, mostra como os dados podem ser capturados com uma qualidade consistentemente desfavorável contra um grupo específico, o que pode gerar diagnósticos menos precisos e, consequentemente, levar a tratamentos inadequados. No fim, ao fornecer esses dados para o treinamento de um sistema baseado em IA, o padrão aprendido pelo modelo pode se configurar em um tratamento discriminatório de viés racial.

Vieses podem ser de difícil identificação, pois o contexto e as restrições em que os dados são gerados podem influenciar a possibilidade de reuso. Neste sentido, outra fonte de risco importante é utilizar a alocação atual de recursos para projetar a alocação futura, pois isso pode perpetuar um tratamento desigual, impedindo um acesso justo a esses recursos. Um exemplo é o caso da Amazon com a sua ferramenta de apoio à seleção de candidatos (DASTIN, 2018). Ela utilizava Inteligência Artificial (IA) para a seleção de

currículos, mas tinha um viés discriminatório contra as mulheres.

Aparentemente, a amostra fornecida estava em sintonia com a população analisada, ou seja, os cargos de tecnologia são, em geral, ocupados por homens. Entretanto, ser mulher reduzia as chances da candidata no processo seletivo, perpetuando uma prática discriminatória que vê os homens como mais aptos às carreiras tecnológicas. É possível que as amostras para treino e teste fossem representantes perfeitas da população-alvo (atuais ocupantes de cargos de tecnologia), o que nos leva à pensar sobre em que circunstâncias o desenvolvimento de modelos deve olhar para além dos algoritmos e dos dados disponíveis, como sugerido por Lipton (2018), "[...] quais são esses outros objetivos e em que circunstâncias eles são buscados?". Neste momento, um objetivo para muitas empresas é buscar uma maior diversidade em suas equipes, logo, essa diversidade deve estar presente nos dados que serão utilizadas para treinar o modelo.

Nos Estados Unidos, um outro exemplo é o de um algoritmo comercial utilizado por duzentos milhões de pacientes que exibia um viés racial significativo. O seu objetivo era o de identificar pessoas que deveriam receber cuidados adicionais. Para identificar a necessidade deste cuidados, utilizava como *proxy* uma previsão dos gastos de saúde. No entanto, os dados utilizados no treinamento capturavam disparidades de acesso entre grupos raciais e isso se refletia em um tratamento que privilegiava pacientes brancos. Os autores afirmam que: "o preconceito surge porque o algoritmo prevê custos de cuidados de saúde em vez de doenças" (OBERMEYER et al., 2019, p. 1). Ainda segundo os autores, "corrigir essa disparidade aumentaria a porcentagem de pacientes negros que recebem ajuda adicional de 17,7% para 46,5%".

Neste caso, depois de concluído o estudo (Obermeyer et al., 2019), os pesquisadores passaram a trabalhar com a empresa para tentar mitigar o viés. Este tipo de desfecho não é comum, pois pesquisadores dificilmente obtêm acesso a algoritmos proprietários ou aos dados utilizados para treinar o modelo, ainda mais quando se trata de saúde, que utiliza dados sensíveis e, em geral, confidenciais (LEDFORD, 2019).

Como último exemplo de vieses que podem ter origem em dados e gerar um tratamento discriminatório, cabe lembrar o caso de Salta (Argentina) em que somente as adolescentes tinham algum papel na prevenção da gravidez precoce (seção 2.3.2). Ou seja, não se pensou em estimar a probabilidade dos adolescentes se tornarem pais nos próximos anos. Toda a responsabilidade é atribuída às meninas. Em outro contexto, o desenho proposto para este problema poderia ser outro.

Em resumo, esta seção apresentou alguns tipos de vieses que se relacionam com o tratamento discriminatório de grupos e indivíduos, sem pretender ser exaustiva na sua cobertura. Neste contexto, em meio a problemas decorrentes dos diversos tipos de vieses e de dados inadequados ou pouco confiáveis, é importante lembrar que os modelos baseados em *Machine Learning* aprendem padrões a partir dos dados fornecidos. Como visto nessa

seção, dados são gerados por dispositivos que podem ter algum grau de imprecisão para grupos distintos (raça, idade, renda, gênero, etc.). Além disso, quando utilizamos dados coletados ao longo do tempo nas mais diversas atividades humanas, é razoável supor que eles registrem os preconceitos e discriminações que estão presentes na sociedade. Consequentemente, reconhecer e agir para não perpetuar esses vieses é fundamental para construirmos um ambiente mais seguro para o uso de IA na sociedade.

3.4 Limitações técnicas e opacidade de modelos de *Machine Learning*

Nos últimos anos, aplicações baseadas em Inteligência Artificial têm sido adotadas com enorme sucesso em áreas como transportes, logística, sistemas de recomendação e tradução de textos. Entretanto, esse êxito tem sido acompanhado pelo aumento da complexidade das soluções, sacrificando o entendimento humano sobre o seu funcionamento. Um dos motivos para a baixa explicabilidade está relacionada, em parte, com o uso crescente de modelos como os baseados em aprendizagem profunda (*Deep Learning*) ou em técnica que permitem a combinação (*ensembles*) de vários modelos em uma solução (Barredo Arrieta et al., 2020).

Em áreas potencialmente sensíveis como a saúde, a falta de transparência é uma limitação técnica que pode ocultar tratamentos discriminatórios (seção 3.3.2) ou, por falta de confiança na solução, funcionar como uma barreira para a adoção da tecnologia. Este último caso, conforme apresentado na seção 3.2, representaria abrir mão de enormes possibilidades.

Nesta seção, parte-se da ideia de que a Inteligência Artificial (IA) pode contribuir enormemente para os desafios atuais de campos como a saúde. No entanto, chama a atenção para as limitações técnicas de algumas arquiteturas de algoritmos que criam soluções pouco transparentes. Essa opacidade impede a compreensão do funcionamento interno dos modelos e, consequentemente, pode tornar pouco consistente o processo de avaliação da qualidade da solução.

Para lidar com essa questão, é fundamental que soluções baseadas em *Machine Learning* (ML) possam, de alguma forma, indicar como as decisões são tomadas, quais dados foram mais relevantes para a decisão ou o que precisa ser alterado para que haja um desfecho diferente. Em resumo, deve-se buscar o nível mais elevado possível de transparência no uso de ML, com o intuito de que esta tecnologia seja aplicada seguindo princípios éticos pactuados.

Para enfrentar os problemas decorrentes da falta de transparência em decisões algorítmicas, um grande esforço tem sido empregado no desenvolvimento da *Explainable*

Artificial Intelligence (XAI), área de pesquisa que busca dar alguma interpretabilidade a modelos de *Machine Learning*. Trata-se de uma área relativamente recente e, por isso, ainda carece de consenso em muitos aspectos, inclusive ao definir conceitos como explicabilidade, interpretabilidade e transparência.

No entanto, já é possível um acordo sobre categorias de modelos que podem ser classificados como mais opacos do que outros, por exemplo, redes neurais profundas são, em geral, menos transparentes do que árvores de decisão. Outro consenso é o de caracterizar esse modelos como "caixas-pretas" por suas estruturas complexas para seres humanos (redes neurais e florestas randômicas, por exemplo) ou por restrições decorrentes de preocupações com a propriedade intelectual.

Além disso, conforme afirma Caruana (2019), "todo conjunto de dados é falho, muitas vezes de maneiras imprevistas e difíceis de detectar", o que é um problema sério já que modelos baseados em *Machine Learning* utilizam os dados para identificar padrões. Caruana (2019) ainda destaca que "se você não consegue entender o que seu modelo aprendeu, então você quase certamente está entregando modelos que são menos precisos do que poderiam ser e que podem até ser arriscados".

Cabe destaque para o fato de que a transparência também pode ser limitada pela complexidade dos dados, mesmo que os modelos utilizados tenham uma arquitetura intrinsecamente transparente. Segundo Lipton (2018, p. 12), "modelos com dimensões suficientemente altas, listas de regras pesadas e árvores de decisão profundas podem ser considerados menos transparentes do que redes neurais comparativamente compactas". Além disso, o autor ainda afirma que, nestas condições, "nem os modelos lineares, os sistemas baseados em regras, nem as árvores de decisão são intrinsecamente interpretáveis". Ou seja, há uma relação entre opacidade e a arquitetura do modelo, mas a complexidade dos dados podem influenciar fortemente a transparência da solução.

De forma simplificada, para lidar com modelos pouco transparentes há basicamente duas abordagens. A primeira é evitá-los, adotando ou desenvolvendo modelos intrinsecamente interpretáveis. A segunda opção é aplicar algum método de análises *post hoc*, o que é feito com o modelo já treinado e buscar interpretabilidade, por exemplo, por meio da identificação de atributos mais relevantes para o desfecho (*feature importance*) ou da seleção de instâncias próximas, mas com desfechos diferentes (explicações contrafactuais). Este tema será discutido em maior profundidade no capítulo 4.

Uma ideia com muito peso na comunidade de *Machine Learning* é a de que modelos mais complexos costumam obter melhores resultados e, consequentemente, isso conduz a uma questão que permeia toda essa discussão: o dilema entre ter maior explicabilidade ou maior acurácia do modelo. No campo da saúde, as decisões devem ser justificadas e a explicabilidade tem um papel importante. Além do mais, nem sempre optar por um modelo intrinsecamente explicável leva a uma perda relevante de acurácia e essa opção

pode evitar problemas graves em decisões de grande impacto.

Neste sentido, Rudin (2019) alerta para o uso de modelos de ML caixa-preta para tomada de decisão crítica, "causando problemas na saúde, justiça criminal e outros domínios". A autora sugere que para essas tarefas sejam projetados modelos intrinsecamente interpretáveis e que seria um mito acreditar que modelos mais complexos são necessariamente mais precisos. Com isso, "pode-se sempre criar um *trade-off* artificial entre precisão e interpretabilidade/explicação, [...] mas isso não é representativo da análise que se apresentaria em um problema real". Além disso, a autora relata um outro efeito da crença na existência desse *trade-off*: muitos pesquisadores são levados a renunciar à tentativa de produzir modelos interpretáveis (RUDIN, 2019, p. 2-3).

Em resumo, o que se busca nesta discussão é destacar o papel central que a explicabilidade/interpretabilidade possui para o desenvolvimento responsável e seguro de aplicações de IA, ainda mais em áreas com decisões críticas como a saúde. Nas seções anteriores (3.2 e 3.3) foram apresentados alguns casos de oportunidades e riscos do uso de IA. Nesta seção, destaca-se a importância da explicabilidade para a adoção de ML, pois sem a compreensão do porquê, para algumas aplicações, um bom desempenho preditivo pode não ser suficiente. Ainda no mesmo tema, o capítulo 4 detalha os principais conceitos e abordagens utilizadas para tentar contornar a principal limitação apresentada nessa seção: a opacidade de modelos de *Machine Learning*. A próxima seção discute a criação de diretrizes de implementação e o papel da regulação como um meio para mitigar riscos e, ao mesmo tempo, tentar criar um ambiente transparente, sem ser uma barreira para a inovação.

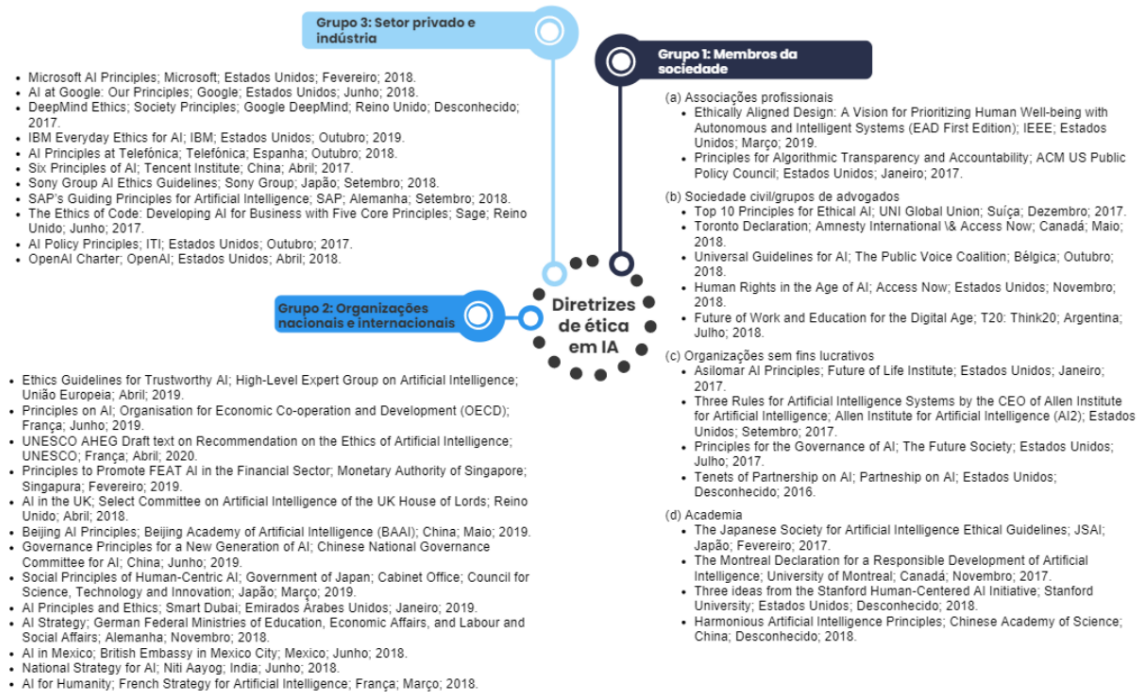
3.5 Esforços regulatórios e diretrizes para a implementação da IA

A Inteligência Artificial (IA) tem sido aplicada com velocidade crescente nas mais diversas áreas. O que tem levado alguns setores da sociedade a expressar preocupação com os efeitos prejudiciais que podem decorrer da tomada de decisão baseada em IA (ver seção 3.3) e a demandar algum nível de regulação estatal. Com isso, várias iniciativas têm sido desenvolvidas por diversos países para mitigar os riscos e fornecer alguma segurança jurídica para as partes interessadas. Infelizmente, não é simples encontrar o correto equilíbrio em que a regulação proteja os interesses de indivíduos e instituições e, ao mesmo tempo, não se torne um empecilho para a inovação. Na seção 3.5.2 são apresentados os principais eixos das propostas de regulação para a IA de alguns dos principais países desenvolvedores.

Reconhecendo a importância de pensar os desafios éticos desde o início, em projetos que incluam algum componente com IA, inúmeras instituições têm se empenhado na construção de documentos com princípios e diretrizes para o uso ético da IA. São iniciativas de governos, empresas privadas e organizações da sociedade civil que buscam algum

consenso em assuntos como transparência, justiça e equidade, privacidade e tratamento não discriminatório. Na figura 9 são listados alguns desses documentos produzidos e tornados públicos entre 2016 e 2020.

Figura 9 – Diretrizes de ética em IA disponíveis publicamente (2016-2020)



Fonte: De Cerqueira, Tives e Canedo (2021, p. 3)

3.5.1 Diretrizes para a implementação da IA

Com o objetivo de analisar como essa discussão tem aparecido em documentos de grande visibilidade e influência, o *Berkman Klein Center for Internet & Society* publicou em 2020 o relatório *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI* (FJELD et al., 2020). Nele são analisados trinta e seis publicações que variam em função do público-alvo, escopo e profundidade.

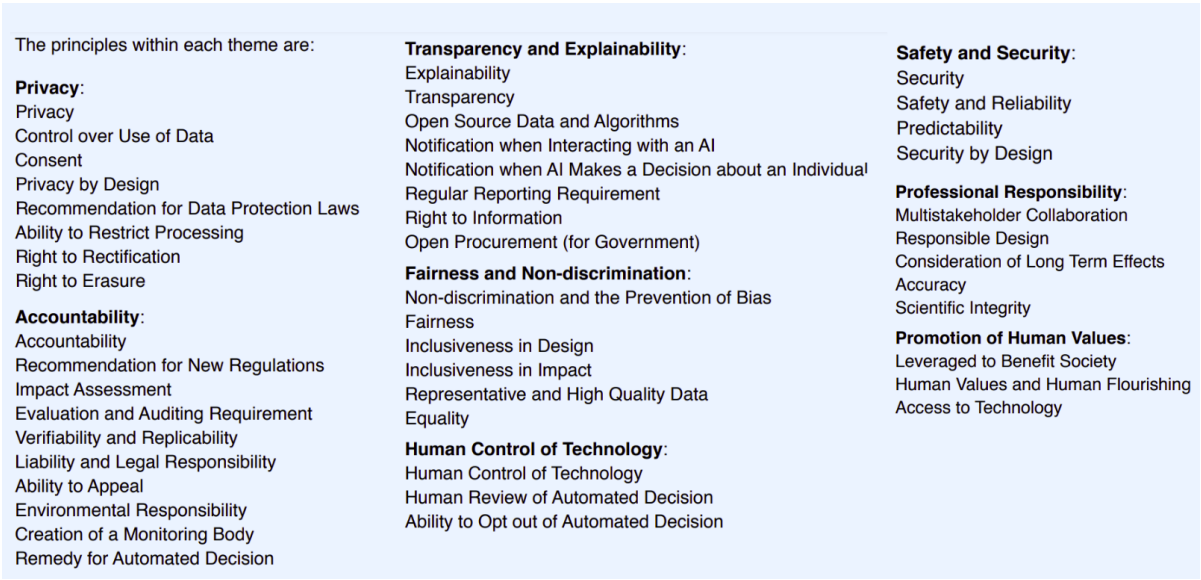
A partir dos textos, são identificados oito temas, cada um deles divididos em princípios. Alguns dos temas são: Responsabilização (*accountability*) e Transparência, Tratamento justo e Não discriminatório, Privacidade e Explicabilidade. Como exemplo, este último tema é dividido em vários princípios, sendo alguns deles: (1) explicabilidade, (2) transparência, (3) dados e algoritmos abertos, (4) direito à informação, (5) aquisições abertas (para governos), (6) notificação ao interagir com uma IA e (7) notificação quando a IA decide sobre um indivíduo. As questões éticas, apesar de serem transversais a todos os temas, encontram maior representatividade no tema sobre Tratamento justo e Não discriminatório (*Fairness and Non-discrimination*).

Um ponto destacado por Fjeld et al. (2020, p. 42) é que o tema Transparência e

Explicabilidade funciona como um pré-requisito para que muitos outros princípios possam ser verificados, o que reforça o seu papel central no desenvolvimento de uma IA que siga princípios éticos, permitindo que o atendimento deste princípios possa ser monitorado e, talvez ainda mais importante para o desenvolvimento da IA, que falhas possam ser identificadas e corrigidas.

A figura 10 apresenta os temas e os princípios identificados por Fjeld et al. (2020).

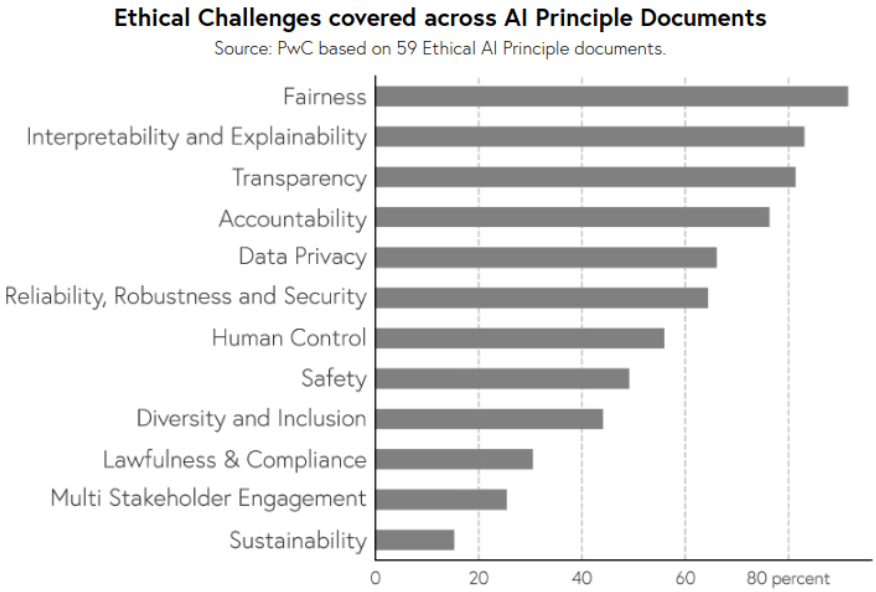
Figura 10 – Temas e princípios para a IA ética



Fonte: Fjeld et al. (2020, p. 7)

Para o amadurecimento do campo de pesquisa, esforços como o de Fjeld et al. (2020, p. 42) na identificação de categorias para o campo de estudos sobre IA e ética têm papel fundamental, pois permitem algum nível de comparação e segmentação do campo. Mas há também outras iniciativas nesta direção, como o *Artificial Intelligence Index 2019 Annual Report da Stanford University* (PERRAULT et al., 2019). Nele são identificados doze desafios éticos mencionados nos documentos analisados. Os autores ainda destacam que a lista não é exaustiva, pois termos como desenvolvimento econômico, redução da pobreza e desigualdade estão ausentes.

Figura 11 – Desafios éticos identificados nos documentos analisados



Fonte: Perrault et al. (2019, p. 150)

Há ainda muito trabalho a ser feito para uma melhor delimitação do campo e o estabelecimento de consensos. Perrault et al. (2019, p. 150) cita uma fala de Anand Rao, líder global de IA da *PricewaterhouseCoopers* (PwC) sobre a falta de maturidade na discussão sobre os princípios éticos na IA.

> A pesquisa em torno da IA ética, especialmente sobre justiça, responsabilidade e transparência [*fairness, accountability, and transparency*] (FAT) de modelos de *Machine Learning* cresceu significativamente nos últimos dois anos. Embora haja um amplo consenso emergindo sobre o conjunto básico de princípios associados à ética e IA, a contextualização desses princípios para setores específicos da indústria e áreas funcionais ainda está em sua infância. Precisamos traduzir esses princípios em políticas, procedimentos e listas de verificação específicos para torná-los realmente úteis e acionáveis para adoção corporativa.

Com isso, podemos ver que ao menos parte da indústria ligada à IA aguarda a definição de forma clara dos princípios éticos a serem seguidos, mas também o estabelecimento de um processo que demonstre a sua real aplicação. Neste sentido, Mittelstadt (2019) destaca que princípios sozinhos não são capazes de garantir uma IA ética, pois, em geral, eles podem ser considerados de alto nível e abstratos para desenvolvimento e implantação de IA e, na prática, fornecem poucas recomendações específicas para traduzir princípios em prática ou mecanismos robustos de responsabilidade legal e profissional.

Sobre a eficácia dessas diretrizes, Hagendorff (2020, p. 10) se pergunta se elas "trazem uma mudança na tomada de decisão individual, independentemente do contexto social mais amplo?". O resultado do estudo relatado pelo autor afirma que "a eficácia

das diretrizes ou códigos de ética é quase nula e que não alteram o comportamento dos profissionais da comunidade tecnológica.” . O autor conclui que o desenvolvimento da IA, na busca pela monetização rápida (e a velocidade é tudo em alguns negócios), não prioriza um enquadramento por uma ética baseada em valores ou princípios, mas obviamente por uma lógica econômica. O autor ainda destaca que “engenheiros e desenvolvedores não são sistematicamente educados sobre questões éticas, nem são capacitados, por exemplo, por estruturas organizacionais, para levantar preocupações éticas” (HAGENDORFF, 2020, p. 10, tradução nossa).

Para tratar essa questão, o estabelecimento de diretrizes e de um arcabouço normativo sobre o desenvolvimento de soluções de IA são fundamentais para resguardar os interesses dos indivíduos, grupos e do próprio Estado. Obviamente, não será fácil alcançar o equilíbrio desejável entre garantir um ambiente que permita a inovação e, ao mesmo tempo, salvaguardar os legítimos interesses envolvidos.

3.5.2 Regulação da implementação da IA

Uma adequada regulação para a Inteligência Artificial (IA) é uma tarefa complexa por vários motivos, inclusive pelo risco de se tornar obsoleta rapidamente devido ao acelerado desenvolvimento do campo. Por outro lado, há uma justa preocupação com os riscos do seu uso para decisões sensíveis e, ao mesmo tempo, muito interesse no promissor potencial para enfrentar os inúmeros desafios em um país de renda média, com tamanha desigualdade e de dimensões continentais, como o Brasil.

É importante destacar que, ao longo de toda esta tese, a cada comentário positivo ou preocupado, é necessária a consciência de que esta discussão deve ser feita sem ignorar que ela se insere em uma enorme disputa por mercados, o que significa dizer que uma regulação muito restritiva levará à desvantagens competitivas e aí reside boa parte da complexidade da regulação: a dificuldade de encontrar um ponto ótimo que minimize os riscos.

Com essa perspectiva em mente, é importante lembrar a previsão do Fórum Econômico Mundial de que até 2030 a tecnologia de IA adicionará mais de US$ 15 trilhões ao produto interno bruto (PIB) global (FIRTH-BUTTERFIELD et al., 2022). Assim, a disputa por cada fatia desse mercado encontra reflexos nas diferentes estratégias que cada país, ou região, tem utilizado para a regulação da IA. Este tema será abordado com mais detalhes nas próximas subseções.

3.5.2.1 Brasil

No Brasil, o estabelecimento de diretrizes e alguma regulação para a IA encontra-se em um estágio inicial. Em abril de 2021 foi estabelecida a Estratégia Brasileira de Inteli-

gência Artificial (EBIA) por meio de uma portaria do Ministério da Ciência, Tecnologia e Inovações (MCTI). Alinhada a diretrizes da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), a EBIA fundamenta-se nos cinco princípios definidos pela Organização para uma gestão responsável dos sistemas de IA. São eles: (i) crescimento inclusivo, o desenvolvimento sustentável e o bem-estar; (ii) valores centrados no ser humano e na equidade; (iii) transparência e explicabilidade; (iv) robustez, segurança e proteção e; (v) a responsabilização ou a prestação de contas (*accountability*) (MCTI, 2021).

No entanto, apesar dos princípios elencados na EBIA estarem alinhados com vários outros documentos publicados nos anos anteriores (SMUHA, 2019; RYAN; STAHL, 2021), alguns autores afirmam que falta "materialidade e um plano de ação mais detalhado para fazer a estratégia sair do papel e se transformar em resultados em benefício da sociedade" (XAVIER, 2021).

Recentemente, com o objetivo de construir um documento que dê maior concretude à discussão e que aponte caminhos mais objetivos para o desenvolvimento da IA no Brasil, o Senado Federal criou uma comissão[9], composta por juristas, para elaborar um projeto de regulação da Inteligência Artificial (IA) no Brasil.

Instituída no dia 30 de março de 2022, ela terá 120 dias para apresentar o resultado final, o que parece ser uma meta ambiciosa, pois dentre outras questões, ela pretende abordar os contextos econômico-sociais e benefícios da IA, desenvolvimento sustentável e bem-estar, inovação, pesquisa e desenvolvimento da IA (fundos de recursos e parcerias público-privadas), segurança pública, agricultura, indústria, serviços digitais, tecnologia da informação e robôs de assistência à saúde[10].

No âmbito público brasileiro, com o objetivo de orientar a adoção e o desenvolvimento das aplicações de IA, uma iniciativa que merece destaque é a Resolução 332 de 21/08/2020 do Conselho Nacional de Justiça (CNJ). Ela dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário. Os capítulos dessa resolução abordam temas como a não discriminação, o respeito aos direitos fundamentais, a necessidade de segurança e o direito do usuário a ser informado sobre o uso de IA nos serviços prestados (CNJ, 2020).

Um ponto importante é que a resolução passa por diretrizes gerais, mas desce a algum nível de especificidade ao definir, por exemplo, que há obrigatoriedade de comunicação ao CNJ sobre o início de pesquisa, desenvolvimento ou implantação de sistemas baseados em IA e que eles devem ser homologados de forma a "identificar se preconceitos ou generalizações influenciaram seu desenvolvimento, acarretando tendências discriminatórias no seu funcionamento" e, caso seja identificado algum viés discriminatório, devem ser

[9] Comissão de Juristas responsável por subsidiar elaboração de substitutivo sobre Inteligência Artificial no Brasil: <https://legis.senado.leg.br/comissoes/comissao?codcol=2504>

[10] <https://www12.senado.leg.br/noticias/materias/2022/03/30/brasil-podera-ter-marco-regulatorio-para-a-inteligenc

adotadas as medidas corretivas cabíveis (CNJ, 2020, p. 5, art. 7º).

O texto ainda destaca que, caso não seja possível eliminar o viés discriminatório, o sistema deve ser descontinuado. Além disso, reconhecendo os riscos envolvidos, a resolução exige uma autorização prévia do CNJ para projetos que envolvam técnicas de reconhecimento facial (CNJ, 2020, p. 9, art. 22º).

Esta tese argumenta que algo semelhante à resolução 332 do CNJ poderia, em parte, contribuir para a construção do arcabouço que regulará a introdução de sistemas baseados em IA no SUS. Além dela, que foi inspirada no pela Carta Europeia de Ética sobre o Uso da Inteligência Artificial em Sistemas Judiciais e seu Ambiente (PRADO; MÜNCH; VILLARROEL, 2022), há muitos outros documentos recentes com objetivos similares e, inclusive, foco no campo da saúde. Nas próximas seções, serão apresentadas as propostas de outros países e regiões para a regulação da IA.

3.5.2.2 Estados Unidos da América (EUA)

Nos Estados Unidos da América (EUA), a responsabilidade pela aprovação dos algoritmos baseados em IA é da *U.S. Food and Drug Administration* (FDA). A partir do conceito de software como um dispositivo médico (*Software as a Medical Device* - SaMD) e da tabela de classificação de riscos, ambos definidos pela *International Medical Device Regulators Forum* (IMDRF), é utilizada uma abordagem que classifica cada SaMD em uma categoria de risco, em função do uso pretendido (US FDA, 2019).

Os dois principais fatores identificados na tabela da IMDRF são: (a) importância das informações fornecidas pelo SaMD para a decisão de saúde; (b) situação ou condição de saúde (figura 12). Assim, os SaMD são classificados em uma das quatro categorias de risco: I, II, III e IV.

Figura 12 – Categorização de risco SaMD IMDRF

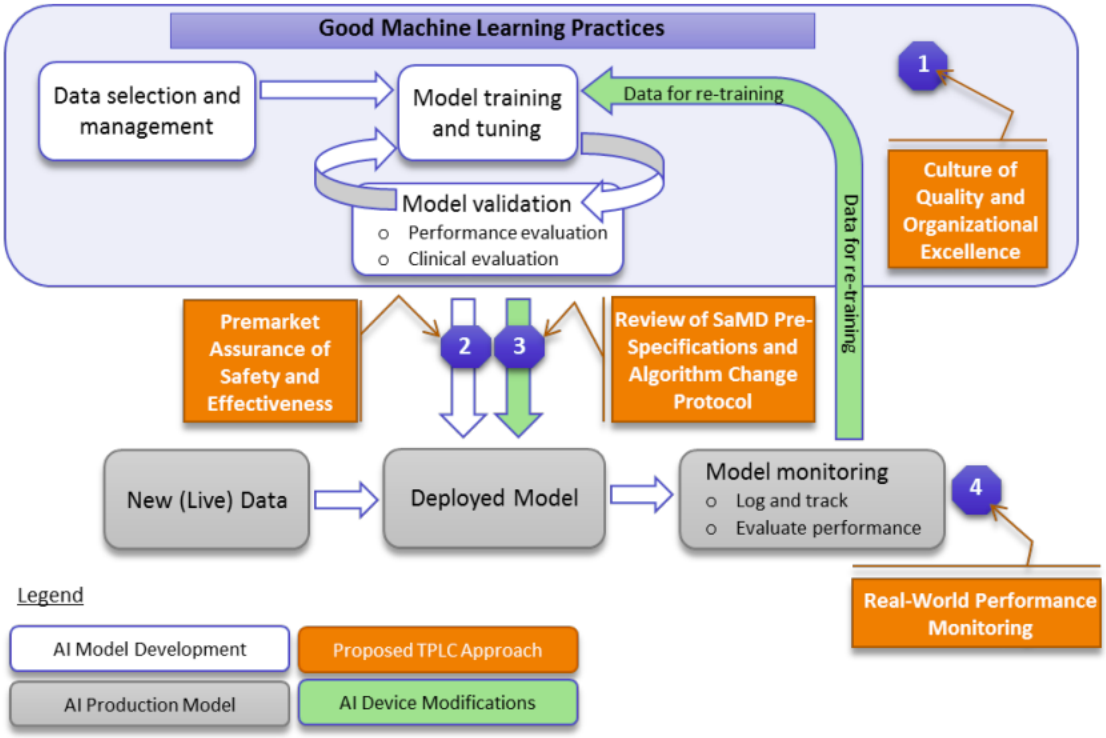
State of healthcare situation or condition	Significance of information provided by SaMD to healthcare decision		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Fonte: US FDA (2019)

Na abordagem da FDA, além da categorização anterior, há também uma diferenciação quanto ao rigor regulatório de soluções que são bloqueadas (*locked*) para aprendizado contínuo e os algoritmos adaptativos. Algoritmos bloqueados fornecem o mesmo resultado

se a entrada não mudar, enquanto os algoritmos adaptativos podem apresentar comportamentos diferentes ao longo do tempo, por exemplo, melhorando o desempenho com novos dados ou quando associado a equipamentos específicos. Na figura 13 é apresentado o ciclo proposto pela FDA para aplicações baseadas em IA, que se aplica aos SaMD baseados em IA/ML que exigem envio pré-mercado. Com isso, ficam isentos os produtos de classe I e II (US FDA, 2019, p. 8).

Figura 13 – Visão geral do fluxo proposto pela FDA para aplicações baseadas em IA



Fonte: US FDA (2019, p. 8)

Nessa abordagem, é avaliada a cultura de qualidade e excelência organizacional da empresa. O objetivo é o de tornar possível avaliar e monitorar um produto de software (SaMD), desde o desenvolvimento pré-mercado até o desempenho pós-mercado, assim como obter uma avaliação contínua da excelência da organização.

Em resumo, os sistemas baseados em IA são avaliados quanto à segurança e eficácia, mas os processos de qualidade do fabricante também são importantes para estabelecer expectativas claras sobre o gerenciamento contínuo dos riscos dos pacientes. Assim, a partir dos relatórios de desempenho do sistema operando no mundo real, o objetivo é fornecer transparência para usuários e à FDA (US FDA, 2019, p. 8-9).

3.5.2.3 União Europeia (UE)

A abordagem da União Europeia (UE) é a de uma regulação baseada no risco da IA. Para isso, são definidos três níveis: i) um risco inaceitável, ii) um risco elevado, iii) um risco baixo ou mínimo.

As aplicações com risco inaceitáveis são proibidas e, dentre outras, incluem aquelas que violem os direitos fundamentais, o uso de técnicas subliminares com potencial significativo para manipular as pessoas e explorar as vulnerabilidades de grupos específicos, como as crianças ou as pessoas com deficiência. Um ponto importante é a garantia para que as pessoas "sejam devidamente informadas e tenham a liberdade de decidir não se sujeitar a uma definição de perfis ou a outras práticas que possam afetar o seu comportamento" (European Commission, 2021, p. 12-13).

Os sistemas de risco elevado são os incluídos em domínios como o de identificação biométrica e categorização de pessoas, gestão e funcionamento de infraestruturas críticas, manutenção da ordem pública, administração da justiça e processos democráticos (European Commission, 2021, anexo III). Para sistemas de risco elevado, há uma lista de requisitos extensa no capítulo 2 do regulamento. O capítulo 3 define as obrigações impostas aos fornecedores e a outras partes envolvidas como importadores, distribuidores, mandatários, etc.

Após essa breve apresentação da estratégia de regulação da IA para a UE, é possível verificar que ela tem como foco o estabelecimento de um arcabouço regulatório mais amplo e com uma definição detalhada de processos e artefatos envolvidos, o que aparentemente pode tornar o processo mais robusto, mas ao mesmo tempo mais lento e custoso, o que destaca a complexidade de encontrar um ponto de equilíbrio adequado na regulação.

Na UE, além da proposta de regulação para a IA, há algumas iniciativas para a disseminação de ferramentas de auto-avaliação como a ALTAI (*The Assessment List on Trustworthy Artificial Intelligence*). Dentre os seus objetivos, está a preocupação em fazer com que as organizações entendam o que é IA confiável, em particular quais riscos um sistema de IA pode gerar e também promover a inovação em IA responsável e sustentável na Europa (High Level Expert Group on AI, 2020).

3.5.2.4 Considerações sobre o desafio da regulação da IA

Há muito a ser feito para garantir um uso ético e seguro de sistemas baseados em IA e, analisando as iniciativas apresentadas nesta seção (3.5.2), fica claro o papel fundamental que a regulação pode desempenhar.

No Brasil, a iniciativa do Conselho Nacional de Justiça (CNJ) com a resolução 332 de 21/08/2020 representa um primeiro passo ao criar um cadastro de projetos baseados em IA, pois isso pode permitir um melhor planejamento na adoção da tecnologia, o

compartilhamento de experiências e evitar o desperdício com esforços redundantes. A resolução também estabelece um processo de homologação, além da definição clara pela descontinuidade de projetos que possam levar a algum tipo de tratamento discriminatório. Um outro ponto de destaque é o reconhecimento da importância da diversidade na composição das equipes para pesquisa, desenvolvimento e implantação da IA.

Abordando um importante requisito para qualificar a efetividade da diversidade, a próxima seção discute o papel da interdisciplinaridade e a necessidade de letramento de todos aqueles que interagem ou são afetados pela IA.

3.6 Interdisciplinaridade e seu papel na busca por uma IA ética

A Ciência de Dados ocupa um papel de destaque para o desenvolvimento de modelos de *Machine Learning* e tem como uma de suas características uma forte interdisciplinaridade. Utilizando a definição adotada pela Plataforma de Ciência de Dados aplicada à Saúde (PCDaS), projeto de pesquisa da Fundação Oswaldo Cruz (Fiocruz):

> Ciência de Dados é um campo de estudo que se destaca pela capacidade de auxiliar a descoberta de informação útil a partir de grandes ou complexas bases de dados, bem como a tomada de decisão orientada por dados. Pode ser definida como um conjunto de estratégias, ferramentas e técnicas para coleta, transformação e análise de dados realizadas por equipes multidisciplinares formadas por pesquisadores com conhecimento substantivo do problema em análise – no nosso caso saúde pública - estatísticos, matemáticos e cientistas da computação (PCDAS, 2021).

A figura 14 apresenta alguns campos fortemente relacionados à Ciência de Dados, mas os desafios a serem enfrentados frequentemente utilizam conhecimentos de muitas outras áreas. Com exemplo, a Ciência de Dados pode recorrer e se beneficiar de discussões e metodologias consolidados em áreas que não costumam figurar nos diagramas que listam as disciplinas que a compõem. Dentre elas, as Ciências Sociais, a Psicologia e a Filosofia, que já discutem temas como ética, moral, explicação e transparência há muito tempo.

Figura 14 – Interdisciplinaridade da Ciência de Dados

Fonte: PCDaS (2021)

Reforçando a importância da interdisciplinaridade, SANDLER e BASL (2019) propõe diretrizes para a criação de comitês de ética de dados e IA. A ideia é que eles possam contribuir para a construção de capacidades de avaliação das implicações éticas de modelos desenvolvidos ou em produção. O autor sugere que a composição desses comitês deve conter especialistas em Ética, Direito, Tecnologia, no tema em análise e de cidadãos.

Ainda sobre interdisciplinaridade, em entrevista ao jornal El País (COLLERA, 2019), Jen Gennai, líder da equipe de inovação responsável na Google, fala sobre iniciativas da empresa para capacitar os seus profissionais para lidar com questões éticas relacionadas a soluções de *Machine Learning*. Segundo Gennai, "nossos engenheiros não estudaram filosofia e não entendem o que significa a ética em seu trabalho. Temos que ajudá-los a internalizar uma série de noções sobre o assunto para que possam aplicá-las ao seu dia a dia".

Na mesma entrevista, Gennai destaca a importância que a ética deveria assumir na formação acadêmica de profissionais envolvidos com *Machine Learning*.

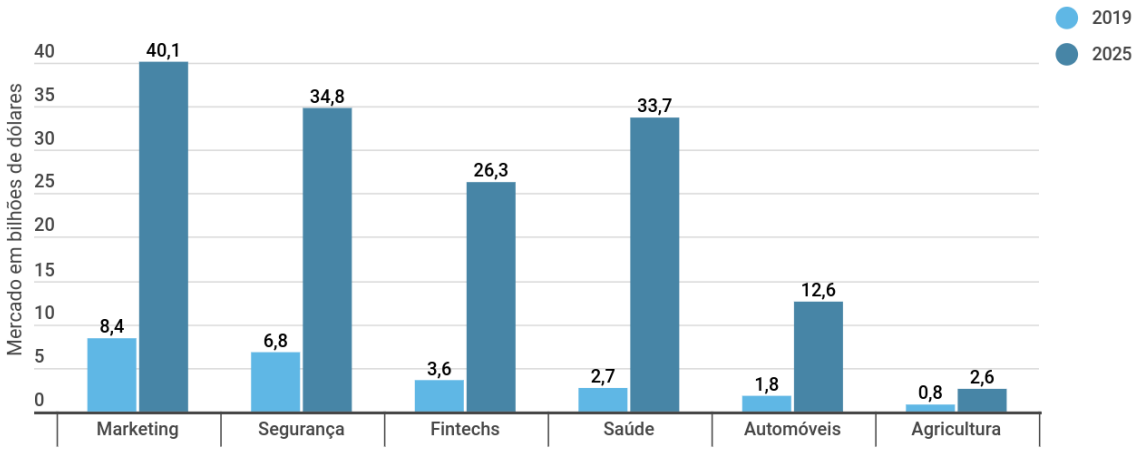
> No meu tempo não se falava de ética na sala de aula, mas agora sabemos o impacto dos avanços da indústria na sociedade e deve ser um assunto central, não um mero acessório em programas tão procurados como os de *machine-learning*. (COLLERA, 2019)

A fala é contundente e chama a atenção para a necessidade de uma discussão mais

ampla na sociedade sobre os desafios éticos associados ao uso de Inteligência Artificial (IA). Nesta tese, reconhecendo o potencial de uma olhar interdisciplinar para mitigar os riscos no uso dessa tecnologia, argumenta-se que esse debate não deve ser restrito às áreas técnicas, pois um número cada vez maior de pessoas será avaliado ou tomará decisões de alto impacto com o apoio de sistemas baseados em IA. Por isso, elas precisam de uma boa compreensão do que é a IA e de como funcionam os seus mecanismos básicos.

Envolver a sociedade em um esforço de alfabetização (*literacy*) sobre a IA é também um ponto destacado em uma publicação do Fórum Econômico Mundial (FIRTH-BUTTERFIELD et al., 2022), dada a importância projetada de crescimento e capilarização da IA na sociedade. Segundo as estimativas apresentadas, o mercado de IA deve atingir um valor de US$ 190 bilhões até 2025 e, até 2030, a tecnologia de IA adicionará US$ 15,7 trilhões ao produto interno bruto (PIB) global.

Figura 15 – Estimativas para o mercado de IA (em bilhões de dólares)



Fonte: adaptado de Singh (2019)

Firth-Butterfield et al. (2022) sugerem que a democratização do acesso à IA, e o seu uso seguro e responsável, passam por promover a alfabetização universal em IA e pela priorização da diversidade no seu desenvolvimento e implantação.

DeCario e Etzioni (2021) descrevem uma pesquisa realizada pelo Allen Institute for AI[11] com 1.547 adultos norte-americanos. A pesquisa aplicou um questionário com 20 perguntas objetivas para avaliar o conhecimento sobre o funcionamento da IA. Apenas 16% dos participantes conseguiram acertar pelo menos 60% das questões, ou seja, o estudo conclui que a grande maioria dos americanos é mal informada sobre o que a IA é capaz de fazer.

> Talvez o analfabetismo em IA não deva nos surpreender. A IA não faz parte dos currículos de nossas escolas, e a principal fonte de informação

[11] Allen Institute for AI: <https://allenai.org/>

> sobre ela hoje, segundo nossa pesquisa, é o YouTube e as redes sociais. No entanto, a IA está transformando o mundo ao nosso redor em um ritmo alarmante; A alfabetização em IA (uma compreensão básica do que pode e do que não pode fazer) é fundamental para informar as decisões cotidianas, adotar políticas econômicas apropriadas e manter nossa segurança nacional. Não estamos defendendo que todos se tornem adeptos da criação de software de IA, mas sim que as pessoas entendam claramente as capacidades, limitações e trajetória da IA e como isso afeta suas vidas diárias (DECARIO; ETZIONI, 2021, tradução nossa)[12].

Para os autores, a alfabetização (*literacy*) em IA assume importância suficiente para estar presente nos currículos escolares (DECARIO; ETZIONI, 2021). Entretanto, isso só seria possível com professores devidamente capacitados no tema.

Tendo o SUS como contexto, é possível supor que o número de gestores ou profissionais de saúde com uma compreensão razoável sobre a IA não seja muito melhor do que os 16% encontrados na pequisa, e isso pode dificultar a construção de um ambiente interdisciplinar de qualidade para a avaliação prévia ou monitoramento de sistemas baseados em IA.

Obviamente, é importante que os usuários do SUS também tenham uma boa compreensão da tecnologia. Logo, torna-se fácil compreender que esse será um esforço de letramento gigantesco e que deve atingir a maior parcela possível da sociedade. Como exemplo deste tipo de esforço, a União Europeia adotou uma iniciativa finlandesa de educação chamada *Elements of AI*[13]. Ela tem como objetivo incentivar o maior número possível de pessoas a descobrir o que é a IA, o que ela permite (e o que não permite) fazer. Segundo o site da iniciativa, mais de 750.000 pessoas já se envolveram nestas capacitações.

Por fim, modelos de *Machine Learning* podem ser aplicados a problemas complexos e com alto impacto sobre a vida de pessoas, grupos, comunidades, populações e instituições. Este impacto nem sempre é claro em um primeiro momento, mas contar com diferentes olhares pode ajudar na identificação e tratamento de questões éticas associadas. Neste sentido, a interdisciplinaridade cumpre um papel importante de duas formas. Seja incluindo profissionais de outras áreas (saúde, humanidades ou jurídicas, por exemplo) nas equipes de desenvolvimento de soluções de IA, seja adicionando à formação profissional conteúdos com debates maduros fora do campo tecnológico.

[12] *Perhaps AI illiteracy shouldn't surprise us. AI is not part of our schools' curricula, and the main source of information about it today, according to our survey, is YouTube and social media. Yet AI is transforming the world around us at an alarming pace; AI literacy (a basic understanding of what it can do and what it cannot do) is critical for informing everyday decisions, adopting appropriate economic policies, and maintaining our national security. We are not advocating that everyone become adept at creating AI software, but rather that people should clearly understand AI's capabilities, limitations, and trajectory and how it affects their daily lives.*

[13] *Elements of AI*: <https://www.elementsofai.com/>

3.7 Considerações sobre este capítulo

A adoção de soluções com componentes baseados em Inteligência Artificial (IA) tem passado por diversos estágios nos últimos anos. Em um primeiro momento, uma euforia justificada pelo sucesso de inúmeras aplicações no comércio, marketing, redes sociais, sistemas de busca, etc. Por isso, era justificável transferir essa expectativa para o campo da saúde (ver seção 3.2) que, além da sua importância para a sociedade, também apresenta uma enorme oportunidade em função dos volumes financeiros envolvidos.

Em um segundo momento, várias questões são apontadas sobre possíveis vieses que geram um tratamento discriminatório sistemático contra pessoas e grupos (ver seção 3.3), o que pode se tornar uma barreira justificável para impedir o uso da IA. Para lidar com isso, é importante reconhecer que há limitações na tecnologia (ver seção 3.4) e, dentre outras, a opacidade é uma das mais importantes, seja pela arquitetura utilizada, seja pelas preocupações quanto à privacidade ou pela defesa de direitos comerciais. Em uma tentativa de trazer mais transparência para este ambiente, muita expectativa tem sido direcionada para a área de pesquisa conhecida como *Explainable Artificial Intelligence* (XAI), que será discutida em maiores detalhes no capítulo 4.

Outro ponto importante, e que tem demandado a atenção especial da sociedade civil, empresas e governos, é a construção de artefatos que contribuam para o desenvolvimento de uma IA comprometida com princípios éticos. São inúmeros documentos com diretrizes (ver figura 9), em geral de adesão voluntária e que, nem sempre são capazes de efetivamente transformar as práticas (MITTELSTADT, 2019).

Nos últimos anos, esforços regulatórios de diversos países (ver seção 3.5) buscam disciplinar a entrada em produção e o monitoramento de sistemas baseados em IA. Na área da saúde, as iniciativas partem da classificação das aplicações em função dos riscos, mas algumas abordagens focam mais no monitoramento da aplicação, com a entrada mais rápida no mercado, e outras no estabelecimento prévio de evidências sobre segurança e eficácia.

Com o objetivo de mitigar riscos, boa parte dos princípios, diretrizes e regulamentos destacam a importância da diversidade na composição das equipes de desenvolvimento, monitoramento e regulação da IA. Entretanto, os ganhos que a diversidade pode trazer serão limitados sem que as pessoas envolvidas tenham uma boa compreensão do potencial e das limitações da tecnologia. Buscando compreender esta questão, a seção 3.6 discutiu a importância da interdisciplinaridade e do letramento (*literacy*) das pessoas que interagem ou são afetadas pela IA.

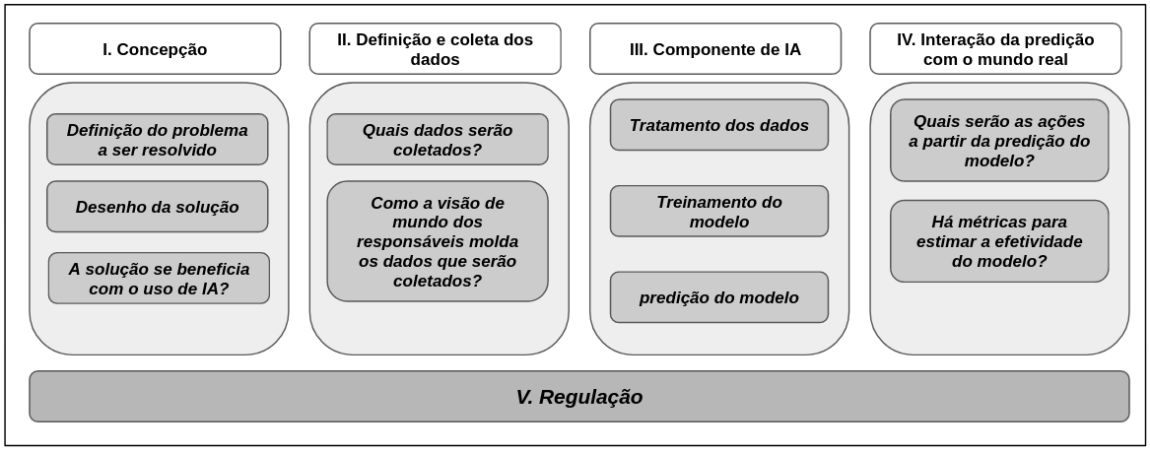
Em resumo, esta pesquisa entende que a IA tem potencial para desempenhar um papel fundamental na ampliação do acesso, na melhoria dos serviços e na otimização da alocação de recursos dos sistemas de saúde. Entretanto, há muito trabalho a ser feito para

desenvolver a confiança na tecnologia. Segundo (MORLEY et al., 2020, p. 8), "os sistemas de saúde não devem ser excessivamente cautelosos com a adoção de soluções de IA-Saúde, mas devem estar atentos aos possíveis impactos éticos para que modelos de governança proporcionais possam ser desenvolvidos".

Inicialmente, para este capítulo, a ideia era entender como equilibrar potenciais, riscos e o papel do crescente campo de pesquisa sobre explicabilidade (XAI), focando na área da saúde. Entretanto, logo ficou claro que a implantação da IA com segurança e eficácia dependia enormemente de fatores não técnicos, como uma compreensão realista do potencial da IA, por todas as partes envolvidas, inclusive gestores, o contexto em que é definido o desenho do projeto e uma compreensão mais profunda sobre os impactos que uma tecnologia pode trazer.

A figura 16 mostra a ideia básica deste capítulo, que entende a solução de IA como apenas um componente de um quadro muito mais amplo, quando consideramos os elementos que podem influenciar o modelo que é construído. Como discutido na seção 2.3.2, uma visão sexista ou racista pode gerar modelos com comportamentos profundamente diferentes e, em grande parte, corrigir esse problema dependerá de um olhar para esse quadro mais amplo, que não está restrito à IA.

Figura 16 – Fluxo simplificado da IA como componente



Fonte: Elaborado pelo autor

Assim, argumenta-se nesta tese que, apesar de todo o seu potencial, a IA é incapaz de corrigir erros de projeto e, além disso, visões enviesadas influenciam os dados a serem coletados e a forma como serão tratados e validados. Consequentemente, modelos treinados com esses dados podem criar limitações no acesso a direitos e serviços para pessoas ou grupos.

O próximo capítulo (4) apresenta a discussão sobre o papel da interpretabilidade/explicabilidade para fornecer uma maior transparência aos modelos de *Machine*

Learning.

4 Interpretabilidade de modelos de *Machine Learning*

Em grande parte, a discussão sobre a interpretabilidade para modelos de *Machine Learning* está relacionada com o fornecimento de alguma compreensão sobre os critérios de tomada de decisão automatizada, o que pode conduzir ao aumento na transparência e, com isso, à construção de uma Inteligência Artificial (IA) eticamente responsável. Neste sentido, a interpretabilidade é um requisito fundamental para atender grande parte das diretrizes que estão sendo propostas por governos, empresas e pela sociedade civil em diversos documentos[1,2,3] para o desenvolvimento e uso da IA.

4.1 Transparência, Interpretabilidade ou Explicabilidade?

Em 2017, a Câmara dos Lordes do Reino Unido apresentou o relatório elaborado por um comitê nomeado para avaliar, dentre outras questões, as implicações éticas e sociais dos avanços na Inteligência Artificial (Parliament UK, 2017, cap. 3). Segundo o relatório, muitos especialistas utilizam o termo transparência, enquanto outros usam "interpretabilidade ou explicabilidade, às vezes indistintamente". Visão semelhante é apresentada por Vilone e Longo (2021b, p. 3) ao dizer que: "a explicabilidade é frequentemente substituída pela noção de interpretabilidade , considerada como sinônimo dentro da comunidade geral de IA".

Em outro exemplo, Doshi-Velez e Kim (2017, p. 3) afirmam que "a *Explainable Artificial Intelligence* (XAI) surgiu como um campo de estudo que foca a pesquisa sobre interpretabilidade do aprendizado de máquina e tem como objetivo fazer uma mudança para uma IA mais transparente", ou seja, os três conceitos podem aparecer juntos, ora de forma intercambiável, ora de forma complementar.

Entretanto, para o desenvolvimento da *Explainable AI* (XAI) como área de pesquisa, ter uma nomenclatura compartilhada para o campo de estudo permitirá o seu desenvolvimento, avaliação e a comparação de trabalhos relacionados (DOSHI-VELEZ; KIM, 2017, p. 9). Com isso, os autores chamam a atenção para a falta de uma definição clara e que esse não é um problema menor.

Lipton (2018), também chama a atenção para este problema, destacando a necessidade de tornar mais claros os conceitos utilizados.

[1] Germany: Artificial Intelligence Strategy
[2] IBM - Everyday Ethics for Artificial Intelligence
[3] Human rights in the age of Artificial Intelligence

> Apesar da falta de uma definição, um crescente corpo de literatura propõe algoritmos supostamente interpretáveis. Disto, você pode concluir que: (1) a definição de interpretabilidade é universalmente aceita, mas ninguém se preocupou em colocá-la por escrito; ou (2) o termo interpretabilidade é mal definido e, portanto, as reivindicações relativas à interpretabilidade de vários modelos exibem um caráter aparentemente científico. Uma investigação da literatura sugere o último (LIPTON, 2018, tradução nossa)

Em seguida, o autor ainda destaca que a interpretabilidade não é um conceito monolítico, "mas várias ideias distintas que devem ser desemaranhadas antes que qualquer progresso possa ser feito".

Apesar de estar fora do escopo desta pesquisa, é fácil reconhecer que é necessário avanço nessa discussão. Assim como para muitos outros autores, utilizaremos os termos interpretabilidade e explicabilidade de forma intercambiável e adotaremos a definição de Doshi-Velez e Kim (2017, p. 2) de que, "no contexto dos sistemas de ML, definimos interpretabilidade como a capacidade de explicar ou apresentar em termos compreensíveis para um ser humano". Essa definição deixa espaço para o entendimento de que os requisitos da interpretabilidade não são únicos, pois públicos diferentes precisarão de explicações em profundidades diferentes. Além disso, verificar a eficácia de uma explicação é uma tarefa complexa, pois "medir se algo foi compreendido ou colocado com clareza é uma tarefa difícil de ser medida objetivamente" (Barredo Arrieta et al., 2020, p. 4).

4.2 A importância da Interpretabilidade

A busca pela interpretabilidade de decisões algorítmicas cumpre um papel importante ao trazer à tona a necessidade de informações que não estão contidas nas previsões e nas métricas calculadas para o modelo. Segundo Lipton (2018):

> Assim, o próprio desejo de uma interpretação sugere que, às vezes, as previsões por si só e as métricas calculadas sobre elas não são suficientes para caracterizar o modelo. Você deve então perguntar: quais são esses outros objetivos e em que circunstâncias eles são buscados? (LIPTON, 2018, tradução nossa)

Um desses objetivos pode ser tentar garantir que o tratamento dado por modelos de *Machine Learning* seja o mais justo possível. Para isso, a interpretabilidade pode agir na identificação de vieses que conduzem a um tratamento discriminatório, por exemplo, restringindo o acesso a serviços de saúde.

Em modelos de *Machine Learning* supervisionados, as medidas de qualidade do modelo (acurácia, precisão, curva ROC, etc.) são derivadas dos dados fornecidos e dos resultados do modelo treinado. No entanto, mesmo quando a amostra fornecida é perfeita-

mente representativa da população, ela pode refletir o tratamento discriminatório presente nas relações sociais intrínsecas aos dados disponíveis.

Uma dificuldade fundamental é que não é simples identificar quais atributos, ou combinação entre eles, podem gerar um efeito discriminatório. Para uma melhor compreensão, adotaremos a definição dada por Calmon et al. (2017) para discriminação.

> Discriminação é o tratamento prejudicial a um indivíduo com base na participação em um grupo legalmente protegido, como raça ou gênero. A discriminação direta ocorre quando atributos protegidos são usados explicitamente na tomada de decisões, também conhecido como tratamento disperso. Mais difundida hoje em dia é a discriminação indireta, na qual atributos protegidos não são usados, mas a confiança em variáveis correlacionadas com eles leva a resultados significativamente diferentes para grupos diferentes. O último fenômeno é denominado impacto disperso. A discriminação indireta pode ser intencional, como na prática histórica de "redlining" nos EUA, em que hipotecas residenciais eram negadas em CEPs [zip codes] habitados principalmente por minorias. (CALMON et al., 2017, p. 1, tradução nossa)

Obviamente, esse efeito pode ter a sua magnitude ampliada em função do contexto de aplicação da IA, como em populações em maior vulnerabilidade. Neste e em muitos outros casos, garantir que as pessoas tenham acesso a explicações sobre como as decisões são tomadas é fundamental para combater os efeitos nocivos de vieses em modelos de ML. Segundo Carvalho, Pereira e Cardoso (2019, p. 10-11), a interpretabilidade traz, dentre outros, o benefício de promover a aceitação social, a segurança dos modelos, a detecção de comportamentos falhos e, para a pesquisa científica, a interpretabilidade pode contribuir para extração do conhecimento capturado pelos modelos.

Por outro lado, como fator complicador, o uso de soluções comerciais de *Machine Learning* gera barreiras adicionais para a transparência das decisões algorítmicas.

Segundo Carvalho, Pereira e Cardoso (2019, p. 8, tradução nossa):

> [...] quando se trata de software proprietário de código fechado, como o COMPAS [4], espera-se que as empresas façam o possível para evitar auditorias e revelar propriedade intelectual. Na verdade, as decisões tendenciosas feitas pelo COMPAS eram mais difíceis de auditar por causa disso.

Casos como o do COMPAS têm chamado a atenção da sociedade civil, de governos e de empresas privadas para a discussão de como implementar uma Inteligência Artificial (IA) que conte com a maior transparência possível e, certamente, parte da solução passa pelo desenvolvimento de tecnologias que garantam algum grau de interpretabilidade sobre o processo de tomada de decisão pelos modelos de ML. Para isso, em geral, as apostas recaem sobre modelos intrinsecamente interpretáveis ou sobre modelos opacos (*black boxes*), mas associados a abordagens de explicabilidade.

[4] <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

4.3 Para quais aplicações a interpretabilidade é importante?

Aplicações com alto impacto na vida de indivíduos e da sociedade, além de setores fortemente regulados, representam nichos em que a interpretabilidade assume relevância quase inquestionável.

Nesse sentido, alguns autores desenham diretrizes para aplicações de ML que caracterizariam situações que podem prescindir de interpretabilidade.

Segundo Doshi-Velez e Kim (2017, p. 3), a explicação não é necessária quando: "(1) não há consequências significativas para resultados inaceitáveis ou (2) o problema é suficientemente bem estudado e validado em aplicações reais em que confiamos na decisão do sistema, mesmo se o sistema não for perfeito". Como exemplos, os autores citam servidores de anúncios, classificação de código postal, sistemas de prevenção de colisão de aeronaves, que tomam decisões sem a intervenção humana.

Infelizmente, não parece ser tão simples delimitar as aplicações que não exigem explicações sobre as suas decisões. Tomemos o caso de servidores de anúncios. Este tipo de aplicação tem como objetivo sugerir produtos ou serviços que podem interessar ao usuário.

Sweeney (2013) analisa exibição de anúncios na página google.com ao fazer buscas por nomes de pessoas. Ela descobriu que nomes atribuídos com maior frequência a pessoas negras nos EUA tinham, naquele momento (2013), maior probabilidade de exibir anúncios sugerindo que alguém com aquele nome tem antecedentes criminais. Por outro lado, a autora ainda demonstra que o mesmo não ocorria quando os nomes eram aqueles frequentemente associados a pessoas brancas.

Nos EUA é proibido discriminar uma pessoa em função de seus antecedentes criminais[5], mas, mesmo que isso não seja verbalizado, a simples sugestão de que uma pessoa possui antecedentes criminais pode ser um inibidor para o acesso a uma vaga de emprego ou a uma promoção, o que pode levar a resultados significativamente prejudiciais, indo de encontro ao item (1) citado por Doshi-Velez e Kim (2017, p. 3).

Com este exemplo, é possível ver parte da complexidade envolvida na tarefa de tentar delimitar tipos de aplicações que podem abrir mão de preocupações sobre a interpretabilidade. Provavelmente, cada situação exigirá uma avaliação individual, que poderá ser revista com o tempo.

A interpretabilidade, além de permitir que a avaliação de como os modelos tomam determinadas decisões, tornando possível a prestação de contas e a verificação da aderência a princípios éticos, também é fundamental para que desenvolvedores possam melhorar, validar e monitorar os modelos. Em muitos casos, não basta que o desempenho preditivo do modelo seja bom, é preciso entender como as decisões são tomadas e se isso acontece

[5] Título VII da Lei dos Direitos Civis de 1964, incluído em 1973

pelos motivos corretos.

Ribeiro, Singh e Guestrin (2016, p. 9) descrevem um experimento com um classificador treinado para identificar se o animal presente em imagens é um lobo ou um husky siberiano. Após testes com um grupo de dez imagens balanceadas, o classificador obtém uma acurácia de 80%. Com uma imagem de lobo e uma de husky siberiano classificadas incorretamente.

Na figura 17.a é exibida uma imagem classificada de forma incorreta. Na imagem 17.b, são destacadas as regiões tidas como mais relevantes pelo classificador, ou seja, um tipo de explicação para a decisão. Segundo os autores, este era um resultado esperado, pois como as imagens utilizadas para treino sempre continham lobos em ambientes com neve, o classificador achou essa uma informação relevante.

No experimento, antes de ter acesso à imagem 17.b, 10 dos 27 participantes confiavam no classificador para uso no mundo real e apensas 12 acreditavam que a neve poderia ser um atributo relevante. Após ver a explicação, somente 3 pessoas continuavam confiando e 25, dos 27 participantes, passaram a reconhecer a neve como um recurso importante.

Figura 17 – Correlação espúria e explicação

(a) Husky classificado como lobo (b) Explicação

Fonte: Adaptado Ribeiro, Singh e Guestrin (2016, p. 9)

No exemplo apresentado, o classificador acertou em 8 das 10 imagens utilizadas no teste, o que pode ter passado a sensação de robustez, já que os testes foram feitos com imagens balanceadas. Entretanto, ao serem apresentados à figura 17.b, tornou-se clara a possibilidade de que os acertos talvez não sejam consequência das características dos animais.

Com isso, justifica-se o argumento central desta seção: a interpretabilidade possui um relevante papel na ampliação da compreensão de elementos de uma predição. O

que, para modelos complexos ou opacos, pode ser determinante na identificação precoce de comportamentos indesejáveis e, assim, torna-se possível evitar o uso de modelos pouco robustos. Logo, a interpretabilidade pode aumentar a transparência para as partes envolvidas e, assim, elevar a confiança no modelo, que é um item fundamental para que *Machine Learning* (ML) possa ser utilizado em situações potencialmente sensíveis.

4.4 Interpretabilidade: taxonomia e tecnologias

Esta seção apresentará alguns conceitos fundamentais relacionados à interpretabilidade. O objetivo é definir a base conceitual que será utilizada nas discussões seguintes. Para isso, será apresentada uma taxonomia com os principais elementos e também algumas soluções tecnológicas relacionadas à *Explainable AI* (XAI).

4.4.1 Taxonomia

A partir de Carvalho, Pereira e Cardoso (2019) e Molnar (2019), podemos classificar os métodos de interpretabilidade de várias formas.

1. **Método específico de modelo ou agnóstico de modelo:** os métodos específicos só podem ser aplicados a um modelo, por exemplo, a interpretação que pode ser feita sobre os pesos de uma regressão linear só se aplica a este modelo. Os métodos agnósticos de modelos são utilizados em modelos já treinados (post hoc). Eles atuam analisando entradas e saídas do modelo, sem acesso à estrutura interna do mesmo;
2. **Método local ou global:** Métodos locais são aplicados para fornecer interpretabilidade para uma instância ou um grupo de instâncias. Métodos globais fornecem informações sobre o comportamento de todo o modelo;
3. **Intrínseco ou *post hoc*:** interpretabilidade intrínseca refere-se a modelos considerados estruturalmente interpretáveis (árvores de decisão, por exemplo). Interpretabilidade *post hoc* refere-se a métodos que são aplicados depois do modelo treinado, buscando explicações que permitam estabelecer relações entre as entradas fornecidas.

Os métodos de interpretabilidade podem ainda ser classificados de acordo com o resultado. Dentre outros, os métodos podem ter como resultado:

1. **Importância do atributo (*feature importance*):** uma lista, ou representação gráfica dela, com o peso dos atributos para um determinado desfecho (método local) ou para o modelo como um todo (método global);
2. **Ponto de dados (*data point*):** os métodos deste tipo retornam instâncias semelhantes, reais ou sintéticas, que possam ser comparadas com a instância em análise.

O objetivo com a comparação pode ser indicar as mudanças mínimas nos atributos que geram um desfecho diferente (explicação contrafactual);

3. **Modelo interno (*internal model*):** são métodos específicos de modelo e que expõem a sua estrutura interna. Para árvores de decisão são retornados atributos e os valores de corte, para modelos lineares, os seus pesos;

4. **Modelo intrinsecamente interpretável:** neste caso, tem-se como resultado um novo modelo (intrinsecamente interpretável) que aproxima o modelo analisado.

4.4.2 Principais tipos de resultados para métodos de interpretabilidade

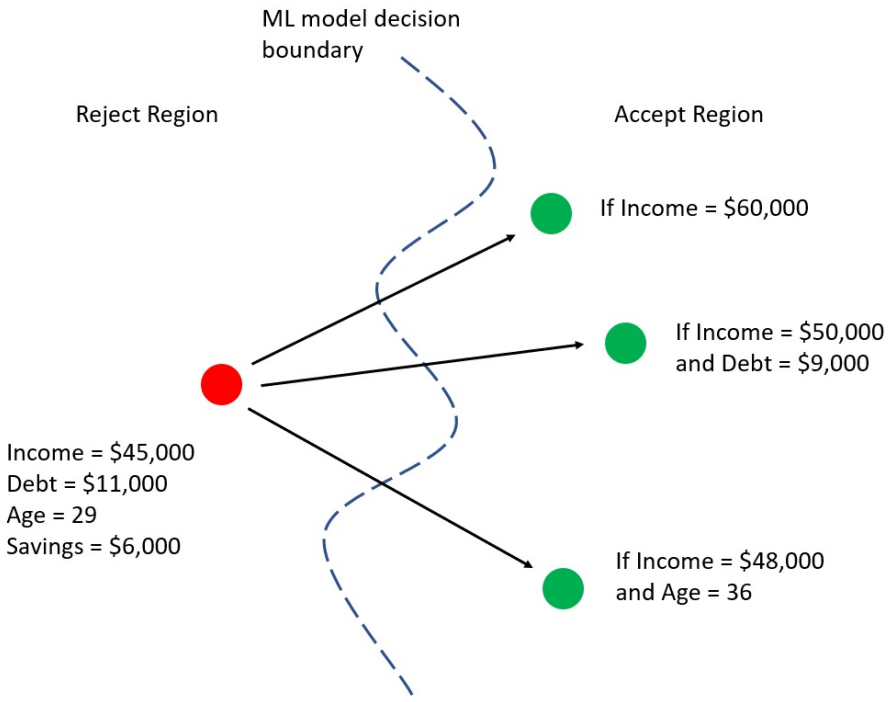
Modelos de *Machine Learning* (ML) tem um enorme potencial para alterar práticas e processos nos mais diversos setores. Por isso, eles precisam ser confiáveis e transparentes, o que é complexo em algumas arquiteturas como as redes neurais profundas. Nesse sentido, a interpretabilidade assume um papel fundamental (ver 4.2), mas os métodos utilizados devem ser robustos e confiáveis ao apresentar uma explicação. Entretanto, Kaur et al. (2020, p. 1), afirmam que "os cientistas de dados confiam demais e fazem uso indevido de ferramentas de interpretabilidade". Com isso em mente, é razoável supor que o desafio pode ser ainda maior para o público não especializado, o que pode se transformar em uma barreira o desenvolvimento da IA em campos como a saúde.

Esta seção discute alguns aspectos de dois dos principais tipos de resultados/saídas dos métodos de interpretabilidade: ponto de dados (*data point*), abordando as explicações contrafactuais, e importância dos atributos (*feature importance*).

4.4.2.1 Explicações contrafactuais

Explicações contrafactuais apresentam instâncias semelhantes (próximas), mas com desfecho diferente. Rudin (2019, p. 19) destaca que explicações contrafactuais "indicam uma mudança nas características que é suficiente (mas não necessária) para que a previsão mude para outra classe". A explicação mostrada para o usuário deve ser a de menor custo (mais simples) para que o usuário possa obter um desfecho diferente.

Figura 18 – Explicação contrafactual



Fonte: PureAI (2020)

No exemplo da Figura 18, são sugeridas três possibilidades. Entretanto, com antecedência, não é simples saber qual mudança será mais fácil para o usuário realizar e isso pode se tornar complicado à medida que o número de atributos aumenta. O que pode ser um problema para a maioria das pessoas, que preferem explicações o mais simples possível.

Todos os métodos possuem vantagens e limitações, e isso não seria diferente com as explicações contrafactuais. Segundo (Molnar, 2019, cap. 6), os métodos contrafactuais não requerem acesso aos dados e ao modelo, apenas acesso à função de predição. De acordo com o autor, "isso é atraente para empresas que são auditadas por terceiros ou que oferecem esclarecimentos aos usuários sem divulgar o modelo ou dados".

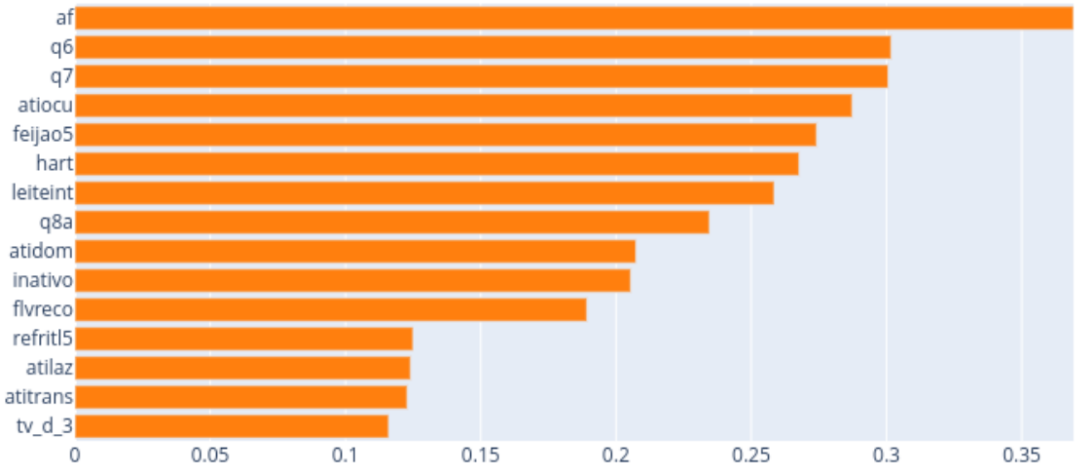
Wachter, Mittelstadt e Russell (2017, p. 6) destacam uma diferença entre a ideia de explicação comumente encontrada na literatura e contrafactuais. A primeira se refere a uma tentativa de transmitir o estado interno ou a lógica de um algoritmo que leva a uma decisão, a segunda descreve uma dependência dos fatos externos que levaram a essa decisão. E, segundo os autores, "esta é uma distinção crucial".

4.4.2.2 Explicações baseadas na importância dos atributos (*feature importance*)

A maior parte dos métodos de interpretabilidade geram algum artefato que exibe a importância dos atributos (*feature importance*) para uma predição (SAARELA; JAUHI-

AINEN, 2021). Um exemplo pode ser visto na figura 19, em que é apresentada uma lista de atributos em ordem decrescente de importância para um modelo de predição de obesidade. Para esse modelo, os três principais atributos para estimar a probabilidade de uma pessoa ser obesa são: a prática de atividade física, a idade e o sexo (atributos af, q6 e q7, respectivamente).

Figura 19 – Predição de obesidade - importância dos atributos



Fonte: Elaborado pelo autor

Conhecer a importância dos atributos em uma decisão pode ser útil para diversos objetivos. Dentre eles, ampliar a compreensão dos dados, aprimorar o modelo e fornecer, para as partes interessadas, alguma interpretabilidade sobre as predições. Além disso, com relação ao escopo, a importância de atributos pode apresentar uma explicação global ou local (ver 4.4.1). Enquanto a explicação global se refere ao modelo como um todo, a explicação local se relaciona a apenas uma entrada específica.

Ressalta-se que, em certas situações, entender quais atributos foram mais relevantes para uma predição é mais importante do que a própria predição. Como exemplo, há situações em que se busca entender quais variáveis foram as mais influentes para um desfecho, o que pode ser útil para direcionar a futura alocação de recursos escassos. Em resumo, explicações podem ajudar a identificar o que pode ser melhorado para evitar um desfecho indesejável.

Por outro lado, é importante lembrar que o valor de uma explicação está ligado à confiança que pode ser depositada nela, o que se torna crítico quando essa explicação pode influenciar a tomada de decisão em ambientes potencialmente sensíveis. No entanto, essa questão aumenta em complexidade quando se verifica que modelos diferentes podem atribuir pesos diferentes para os atributos envolvidos. Segundo Saarela e Jauhiainen (2021, p. 2, tradução nossa):

> As explicações mais comuns para os modelos de classificação são as importâncias dos atributos. [...] Mais precisamente, nos referimos à importância do atributo como uma medida da contribuição individual do atributo correspondente para um classificador particular, independentemente da forma (por exemplo, relação linear ou não linear) ou direção do efeito da característica. Isso significa que as importâncias dos atributos dos dados de entrada dependem do modelo de classificação correspondente e que um recurso importante para um modelo pode não ser importante para outro modelo. [6]

Também nesse sentido, Fisher, Rudin e Dominici (2019) afirmam que a importância dos atributos pode variar entre modelos diferentes. Para lidar com essa questão, os autores propõem uma abordagem (*model class reliance* - MCR) em que a importância dos atributos é estimada como um intervalo, a partir dos valores atribuídos por um conjunto de modelos diferentes, mas com desempenho similar e satisfatório. Segundo Fisher, Rudin e Dominici (2019) , "[...] o MCR fornece uma medida de importância mais abrangente e robusta do que as medidas tradicionais de importância para um único modelo".

Neste ponto fica clara uma questão importante: o processo de busca por interpretabilidade é derivado, em grande parte, da opacidade de algumas soluções de ML. E a cada fragilidade detectada nas abordagens que buscam trazer interpretabilidade, mais complexidade tem sido adicionada, em geral, aumentando enormemente o esforço computacional para obter alguma saída mais confiável.

A questão apresentada no parágrafo anterior é discutida em mais detalhes nas seções 4.7 e 4.8.

4.5 Robustez e estabilidade das explicações

Robustez e estabilidade são dois conceitos importantes para a confiança em uma explicação. Segundo Vilone e Longo (2021b, p. 4), robustez está relacionada à resistência de um método de explicabilidade para suportar, sem alterar a previsão do modelo, pequenas perturbações nos atributos de entrada. Por outro lado, estabilidade é a propriedade de manter-se consistente, fornecendo explicações semelhantes para entradas similares.

Babic et al. (2021) faz uma distinção entre IA explicável e interpretável. Esse último tipo se refere a soluções intrinsecamente interpretáveis (árvores de decisão, regressão linear, etc.), enquanto IA explicável se relaciona com modelos opacos (*black box*) que obtém alguma explicabilidade por meio de modelos que aproximam o seu funcionamento por uma

[6] *"The most common explanations for classification models are feature importances. [...] More precisely, we refer to feature importance as a measure of the individual contribution of the corresponding feature for a particular classifier, regardless of the shape (e.g., linear or nonlinear relationship) or direction of the feature effect. This means that the feature importances of the input data depend on the corresponding classification model and that a feature important for one model may be unimportant for another model"*

função transparente. Essa aproximação é, em geral, imperfeita e local, o que pode gerar problemas para que o modelo seja confiável.

> Para que um algoritmo explicável seja confiável, ele precisa exibir alguma robustez. Com isso, queremos dizer que o algoritmo de explicabilidade deve normalmente gerar explicações semelhantes para entradas semelhantes. No entanto, para uma mudança muito pequena na entrada (por exemplo, em alguns pixels de uma imagem), um algoritmo de IA/ML explicável aproximado pode produzir explicações muito diferentes e possivelmente concorrentes, com tais diferenças não sendo necessariamente justificáveis ou compreendidas mesmo por especialistas. Um médico usando um dispositivo médico baseado em IA/ML naturalmente questionaria esse algoritmo[7] (BABIC et al., 2021, p. 2, tradução nossa)

Logo, é razoável esperar que os modelos baseados em IA, assim como as explicações apresentadas, sejam confiáveis, robustas e estáveis, pois esses são requisitos importantes para identificar falhas, atribuir responsabilidade e, consequentemente, aprimorar os modelos. É importante destacar que, segundo Habli, Lawton e Porter (2020), ainda não existe consenso sobre como atribuir responsabilidades (*accountability*) quando decisões são tomadas ou apoiadas por uma ferramenta baseada em IA. Esse elemento, associado com falta de robustez introduz fragilidade no processo, que pode se transformar em uma barreira à adoção da tecnologia, pois não seria razoável utilizar um modelo que apresenta explicações ou predições contrastantes para entradas similares, sem que alguma justificativa razoável, e compreensível por humanos, seja apresentada.

> Argumenta-se frequentemente que a IA/ML explicável dá suporte a responsabilidade [*accountablitity*] algorítmica. Se o sistema cometer um erro, pensa-se, será mais fácil refazer nossos passos e delinear o que levou ao erro e quem é o responsável. Embora isso geralmente seja verdade para sistemas de IA/ML interpretáveis, que são transparentes por *design*, o mesmo não é verdade para sistemas de IA/ML explicáveis, porque as explicações são raciocínios *post hoc*, que apenas aproximam imperfeitamente a função real que levou à decisão. [...] Assim, vincular a explicabilidade à responsabilidade pode revelar-se um engano[8] (BABIC et al., 2021, p. 2, tradução nossa)

Cabe destaque ainda para o fato de que os sistemas modernos baseados em IA sejam, em geral, uma composição de vários componentes, cada um deles podendo ser uma

[7] *For an explainable algorithm to be trusted, it needs to exhibit some robustness. By this, we mean that the explainability algorithm should ordinarily generate similar explanations for similar inputs. However, for a very small change in input (for example, in a few pixels of an image), an approximating explainable AI/ML algorithm might produce very different and possibly competing explanations, with such differences not being necessarily justifiable or understood even by experts. A doctor using such an AI/ML-based medical device would naturally question that algorithm.*

[8] *It is often argued that explainable AI/ML supports algorithmic accountability. If the system makes a mistake, the thought goes, it will be easier to retrace our steps and delineate what led to the mistake and who is responsible. Although this is generally true of interpretable AI/ML systems, which are transparent by design, it is not true of explainable AI/ML systems because the explanations are post hoc rationales, which only imperfectly approximate the actual function that drove the decision. [...] Thus, linking explainability to accountability may prove to be a red herring.*

black box, com explicações nem sempre robustas (BABIC et al., 2021, p. 2). Assim, a explicação final do sistema de IA estará sujeita à combinação de todas essas imprecisões.

Discutindo algumas abordagens XAI, diversos trabalhos apontam a possibilidade de manipulação da explicação, mesmo sem alteração do desfecho, o que demonstra fragilidade e, obviamente, diminui a confiança no método (SLACK et al., 2021). Dombrowski et al. (2019) apresenta em seu trabalho um algoritmo que permite manipular uma imagem, de forma geralmente imperceptível, e obter um mapa de explicação (*explanation map*) arbitrário. Na figura 20 são apresentadas duas imagens (original e manipulada) e os seus mapas de explicação, que apesar das evidentes diferenças, mantém a mesma classificação da imagem pelo modelo.

Figura 20 – Manipulação do mapa de explicação (*explanation map*)

Fonte: Dombrowski et al. (2019, p. 1)

A manipulação das explicações levanta uma série de preocupações. Uma delas está relacionada à solidez das informações extraídas dessas explicações. Segundo Dombrowski et al. (2022, p. 1), "como algumas explicações são suscetíveis até mesmo a perturbações de entrada aleatórias, parece questionável se muito *insight* pode ser derivado da inspeção de tais explicações". Além disso, quando as explicações deixam de ser evidências confiáveis, verificar se um tratamento gera algum dano indevido é ainda mais desafiador. E essa é uma fragilidade que pode ser explorada por agentes de má fé em uma prática é conhecida com *fairwashing*.

Em situações de *fairwashing*, um fornecedor pode manipular as explicações para promover a falsa percepção de que um modelo de ML respeita alguns valores éticos (AÏVODJI et al., 2019). Neste sentido, Aïvodji et al. (2021, p. 1) afirma que "esse ataque pode afetar significativamente os indivíduos que receberam um resultado negativo seguindo

a previsão do modelo, privando-os da possibilidade de contestá-lo".

Felizmente, muitos trabalhos têm se dedicado a propor abordagens que aumentem a robustez de alguns métodos que fornecem explicações, mas talvez ainda não seja possível afirmar que eles estejam maduros (ALVAREZ-MELIS; JAAKKOLA, 2018; ANDERS et al., 2020).

Por fim, quando o foco é o uso de soluções baseadas em IA em ambientes potencialmente sensíveis, os exemplos apresentados nesta seção destacam a importância do estabelecimento de processos que assegurem os mais altos níveis de qualidade para os modelos, mas também para os métodos de explicação. Como alternativa a modelos opacos, que necessitam de métodos de interpretabilidade, há uma aposta no desenvolvimento e uso de modelos intrinsecamente interpretáveis (ver seção 4.7).

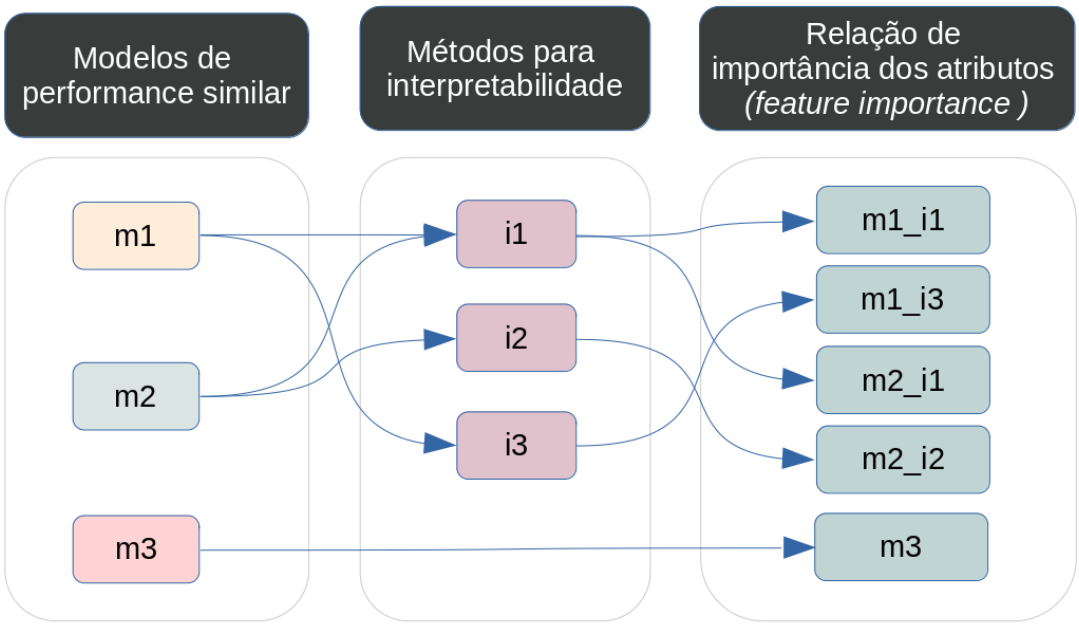
4.6 Multiplicidade preditiva e explicações discrepantes

Diversos trabalhos, alguns deles citados ao longo desta seção, mostram que a partir de um único conjunto de dados é possível treinar modelos diferentes com performance similar e bom desempenho. Segundo Marx, Calmon e Ustun (2020), "problemas de previsão muitas vezes admitem modelos concorrentes que têm um desempenho quase igualmente bom". Com isso, surge a questão: se não há diferença razoável no desempenho, qual modelo deve ser utilizado?

Por outro lado, métodos diferentes de interpretabilidade podem apresentar listas divergentes de importância dos atributos (ver 4.4.2.2). Segundo Fisher, Rudin e Dominici (2019, p. 2), nesses casos o "modelo usado por um analista pode contar com informações de covariáveis totalmente diferentes do modelo usado por outro analista".

A figura 21 ilustra essa questão. Nela é apresentado um suposto cenário com modelos de desempenho similar $m1$, $m2$ e $m3$, sendo este último, por hipótese, intrinsecamente transparente e, por isso, não necessitaria da aplicação de um método de interpretabilidade após o treinamento. Como resultado final, cinco listas de importâncias dos atributos diferentes seriam geradas, o que pode levar a divergência entre elas e, assim, tornar frágeis as afirmações sobre quais são realmente os atributos nos quais se basearam as decisões.

Figura 21 – Modelos similares e possíveis diferenças na importância dos atributos



Fonte: Elaborado pelo autor

A discussão apresentada aqui é proposta para destacar que, quando olhamos para um conjunto de pessoas afetadas pelas decisões tomadas ou apoiadas por modelos de *Machine Learning*, as métricas de qualidade do modelo podem não ser o suficiente para escolher o modelo mais adequado. No exemplo apresentado na figura 22, é ilustrada a ideia de que é possível ter um conjunto de modelos similares ($m1$, $m2$ e $m3$), cada um deles mapeando corretamente uma região (neste exemplo, 80%) das entradas fornecidas em um conjunto de validação do modelo.

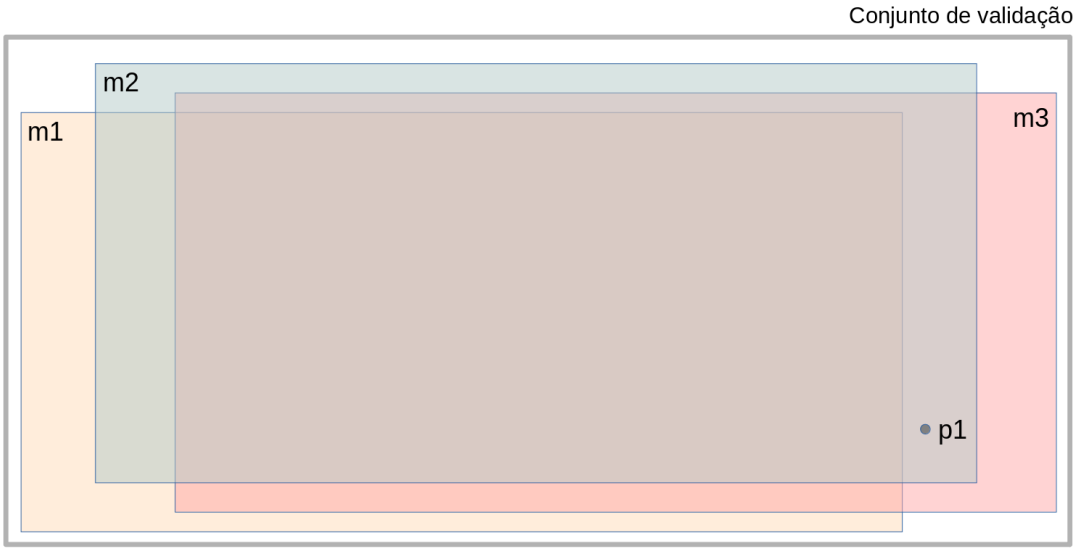
Entretanto, apesar dos três modelos serem similares, para a pessoa $p1$, a troca de $m2$ ou $m3$ por $m1$ fará com que ela seja classificada da forma errada ou desfavorável, ao mesmo tempo que outras pessoas passam a ser classificadas de forma diversa. Esse efeito, em que modelos concorrentes atribuem previsões conflitantes foi definido por Marx, Calmon e Ustun (2020, p. 1-2) como multiplicidade preditiva. Os autores propõem medidas formais (ambiguidade e discrepância) para medir a gravidade da multiplicidade preditiva em problemas de classificação.

Provavelmente, abordagens como essa devam começar a fazer parte das avaliações necessárias antes da adoção de soluções baseadas em IA, ainda mais em campos com decisões de alto risco. Segundo Marx, Calmon e Ustun (2020, p. 8, tradução nossa):

> A multiplicidade de previsões pode mudar a forma como construímos e implantamos modelos em aplicações voltadas para o ser humano. Em tais ambientes, apresentar às partes interessadas informações significativas sobre a multiplicidade preditiva pode levá-las a pensar cuidadosamente

> sobre qual modelo implantar, considere a possibilidade de atribuir previsões favoráveis a indivíduos que recebem previsões conflitantes, ou renunciar totalmente à implantação[9]

Figura 22 – Modelos de mesmo desempenho



Fonte: Elaborado pelo autor

A proposta desta seção foi discutir o fato de que decisões e explicações de *Machine Learning* podem ser discrepantes. Elas podem variar por um enorme número de fatores, mas aqui nos focamos na questão dos modelos de ML e nos métodos de explicação. Esse tópico compõe o panorama alvo desta tese, que se propõe a mapear componentes críticos para a adoção de IA pelo Sistema Único de Saúde (SUS), tendo como preocupação, principalmente, o princípio da Equidade.

4.7 A opção por modelos transparentes

Ao longo deste capítulo, boa parte do texto tem como foco a importância, as oportunidades e os desafios de fornecer interpretabilidade para modelos opacos. O campo conhecido como *Explainable Artificial Intelligence* (XAI) encontra-se em acelerado desenvolvimento, com novos métodos sendo propostos (ZHOU; RIBEIRO; SHAH, 2022; AGARWAL et al., 2020), assim como muito trabalho para tornar mais robustas as soluções disponíveis (AGARWAL et al., 2022; RONG et al., 2022).

Por outro lado, alguns autores afirmam que, sempre que possível, a opção deveria ser por modelos intrinsecamente transparentes, principalmente em situações que envolvam

[9] *Reporting predictive multiplicity can change how we build and deploy models in human-facing applications. In such settings, presenting stakeholders with meaningful information about predictive multiplicity may lead them to think carefully about which model to deploy, consider assigning favorable predictions to individuals who receive conflicting predictions, or forgo deployment entirely.*

decisões de alto risco, muito comuns em campos como o da saúde. Segundo Rudin (2019, p. 1, tradução nossa):

> Em vez de tentar criar modelos que sejam intrinsecamente interpretáveis, houve uma explosão recente de trabalho sobre 'ML explicável', onde um segundo modelo (*post hoc*) é criado para explicar o primeiro modelo de caixa preta. Isso é problemático. As explicações geralmente não são confiáveis e podem ser enganosas, como discutiremos a seguir. Se, em vez disso, usarmos modelos que são intrinsecamente interpretáveis, eles fornecem suas próprias explicações, que são fiéis ao que o modelo realmente calcula[10]

A autora ainda destaca que a opção por modelos opacos (*black boxes*) está relacionada ao fato de haver uma crença de que modelos mais complexos são também mais precisos. No entanto, isso não seria verdade, especialmente para dados estruturados em que os atributos são significativos. Por outro lado, a autora argumenta que *black boxes* podem ser úteis em decisões de alto risco, mas atuando como um componente no processo de descoberta do conhecimento (RUDIN, 2019, p. 2, 19).

Obviamente, não há motivos para que os modelos intrinsecamente interpretáveis fiquem restritos aos algoritmos de regressão linear e logística, árvore de decisão e etc. Assim, Nori et al. (2019) propõem *Explainable Boosting Machine* (EBM), um modelo interpretável (*glassbox*) com acurácia comparável a soluções do tipo *black box*. Na figura 23 é apresentada uma comparação de desempenho entre alguns dos principais algoritmos do tipo *black box* e a EBM, o que destaca o potencial da proposta.

Figura 23 – Desempenho de classificação para modelos em conjuntos de dados (linhas, colunas)

Classification Performance (AUROC)					
Model	heart-disease (303, 13)	breast-cancer (569, 30)	telecom-churn (7043, 19)	adult-income (32561, 14)	credit-fraud (284807, 30)
EBM	**0.916**	**0.995**	**0.851**	**0.928**	0.975
LightGBM	0.864	0.992	0.835	**0.928**	0.685
Logistic Regression	0.895	**0.995**	0.804	0.907	0.979
Random Forest	0.89	0.992	0.824	0.903	0.95
XGBoost	0.87	**0.995**	0.85	0.922	**0.981**

Fonte: Nori et al. (2019, p. 4)

Mesmo para aplicações de visão computacional, em que redes neurais profundas possuem um papel de destaque, há propostas de algoritmos interpretáveis (CHEN et al.,

[10] *Rather than trying to create models that are inherently interpretable, there has been a recent explosion of work on "Explainable ML," where a second (posthoc) model is created to explain the first black box model. This is problematic. Explanations are often not reliable, and can be misleading, as we discuss below. If we instead use models that are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes*

2019). No entanto, Rudin (2019, p. 12) afirma que um desafio para o desenvolvimento de modelos interpretáveis para visão computacional é a falta de uma definição clara de interpretabilidade específica no domínio e que, uma vez estabelecida essa definição, seria possível incorporá-la ao algoritmo.

A questão central quando se advoga pelo uso, sempre que possível, de modelos interpretáveis, está na impossibilidade de saber quais atributos e pesos estão sendo considerados em uma decisão de um modelo opaco, o que pode conduzir a danos difíceis de serem rastreados e mitigados.

Como alerta para essa questão, Caruana et al. (2015, p. 1) descreve um estudo de risco de óbito por pneumonia em que foi considerado arriscado utilizar um modelo opaco (rede neural) e, em substituição, optou-se por um modelo transparente (regressão logística), mas menos preciso. A decisão foi tomada após um modelo baseado em regras encontrar uma regra que associava pacientes com pneumonia, com histórico de asma, a um menor risco de morrer de pneumonia do que a população em geral, o que poderia diminuir a prioridade de atendimento desses pacientes, aumentando a probabilidade de óbito. Na verdade, a associação entre asma e baixo risco de óbito estava realmente nos dados, mas era fruto do fato de que esses pacientes, quando chegavam no hospital, não eram simplesmente internados, mas iam diretamente para a UTI (Unidade de Tratamento Intensivo). Este tratamento diferenciado, não capturado pelos dados, diminuía a probabilidade de óbito.

É importante a compreensão de que existem eventos não registrados diretamente nos dados (internação imediata em UTI para pacientes com asma e pneumonia), mas que suas consequências podem atuar para gerar privilégios ou danos às partes envolvidas (predição de menor risco de óbito para esses pacientes). Para Caruana et al. (2015, p. 2), encontrar essa associação equivocada foi fundamental para optar por uma abordagem transparente, pois seria difícil, pela falta de a inteligibilidade, saber quais outros problemas também precisariam de correção.

Partindo desse exemplo e de uma situação hipotética em que existam dois grupos de modelos disponíveis (opacos e transparentes), mas em que os opacos sempre têm melhor desempenho, duas questões se colocam em discussão: (1) qual é a perda aceitável em desempenho para descartar o uso de um modelo opaco (*black box*)? E, por outro lado, (2) quão melhor deve ser um modelo opaco para que se abra mão de modelos transparentes?

Embora sejam questões semelhantes, elas partem de pontos opostos. Por isso, para situação potencialmente sensíveis, argumentamos que a escolha e o descarte de modelos deve ser documentada de forma que fique clara a motivação, as restrições e os limites de desempenho estabelecidos. Ao mesmo tempo, após a definição dos limites aceitáveis de performance, também argumentamos que a preferência deva ser por modelos interpretáveis, que estejam dentro destes limites.

Além disso, outra questão importante se relaciona à discussão sobre a necessidade de que o uso de IA siga princípios éticos de justiça e equidade, mas nem sempre as definições necessárias estão claras e pactuadas. Segundo Rudin, Wang e Coker (2020, p. 6), "não importa qual definição técnica de justiça seja escolhida, é mais fácil debater a justiça de um modelo transparente do que um modelo proprietário", o que mais uma vez ressalta o papel da interpretabilidade para a construção de um ambiente seguro para o uso de IA.

Apesar disso, modelos opacos continuam sendo utilizados, mesmo quando é possível substituí-los por soluções mais simples e transparentes. Argumenta-se que proteger a propriedade intelectual, permitindo o uso de modelos opacos, incentiva as empresas a realizar pesquisa e desenvolvimento (RUDIN; WANG; COKER, 2020, p. 6). Por outro lado, é possível que o efeito seja tornar desinteressante o desenvolvimento de novas abordagens mais transparentes e, possivelmente, de menor custo. Este é um ponto relevante quando se olha para o panorama de gastos crescentes nos sistemas de saúde. Nesse ambiente, a regulação pode assumir um papel crucial, criando um cenário que priorize modelos transparentes, o que é fundamental para o campo da saúde e para o SUS, garantindo o respeito ao princípio da Equidade.

4.8 Considerações sobre este capítulo

Este capítulo teve como foco a importância da interpretabilidade para um uso seguro de soluções baseadas em ML, para a responsabilização (*accountability*) e, em geral, para a compreensão dos padrões identificados pelos modelos nos dados de treinamento (seções 4.1 e 4.2).

Em seguida, foram discutidas algumas situações em que são necessárias explicações sobre o funcionamento dos algoritmos, pois nem sempre a predição combinada com uma boa acurácia (ou qualquer outra métrica), é suficiente para garantir que o modelo tem um funcionamento correto (seção 4.3). Nesta seção, argumenta-se que não é simples a classificação de quais tarefas exigirão interpretabilidade, pois a compreensão das partes envolvidas sobre o problema pode mudar ao longo do tempo, além da possibilidade de identificação tardia dos riscos envolvidos.

Na seção 4.4 foi apresentada uma breve taxonomia e os principais conceitos envolvidos no ambiente de *Explainable Artificial Intelligence* (XAI), além dos dois dos principais resultados/saídas dos métodos de interpretabilidade (explicações contrafactuais e importância dos atributos).

Atribui-se aos métodos de interpretabilidade o papel de demonstrar a corretude e adequação de modelos opacos e, com isso, tonar possível a responsabilização (*accountability*) em ambientes com o uso de componentes baseados em IA. A seção 4.5 apresentou alguns dos principais requisitos para tornar as explicações confiáveis e, ao mesmo tempo,

faz um contraponto com os modelos intrinsecamente interpretáveis, que por *design* são transparentes. Foram apresentados trabalhos que mostram a possibilidade de manipulação das explicações, o que pode favorecer a prática conhecida como *fairwashing*.

Na seção 4.6 é discutido o fato de que é possível encontrar, dada uma tarefa e um conjunto de dados, modelos de desempenho similar e de boa performance, sem que eles sejam coincidentes, ou mesmo que um seja subconjunto do outro. Em resumo, isso significa que uma mesma pessoa pode ter tratamentos diferentes em função do modelo selecionado e que essa é uma questão relevante quando a decisão envolvida envolver um alto risco. Esse caso, quando combinado com o fato de que métodos de interpretabilidade diferentes podem gerar explicações substancialmente diferentes, torna este um tema relevante quando se pensa em estimar a adequação de um modelo a ser colocado em produção.

Modelos opacos (*black box*) são frequentemente reconhecidos pelo bom desempenho. Por outro lado, àqueles reconhecidos como intrinsecamente interpretáveis, é atribuído um desempenho inferior. A seção 4.7 discutiu essa questão a partir de alguns trabalhos recentes e de novos algoritmos propostos. Eles demonstram que em muitas situações é viável obter modelos transparentes tão precisos como as *black boxes*. Argumenta-se que, sempre que possível, a opção deve ser por modelos interpretáveis, especialmente em decisões de alto risco. No entanto, isso não significaria abandonar os modelos opacos, pois eles continuariam a desempenhar um papel para auxiliar a formulação de hipóteses, além do estabelecimento de uma linha de base para o desempenho de modelos transparentes.

Este capítulo completa o anterior (capítulo 3), que discutiu um quadro mais amplo do ambiente em que se desenvolvem as aplicações de IA. Lá, o foco eram as oportunidades, riscos, vieses, esforços regulatórios e outros assuntos correlatos. Aqui, mais circunscrito às reais possibilidades e limitações de modelos de *Machine Learning*, o foco se dirige para as limitações dos métodos de interpretabilidade, assim como os possíveis impactos em ambientes sensíveis e de alto risco como o setor saúde.

5 Experimento: multiplicidade preditiva e consistência de explicações por meio de listas de importância dos atributos

Neste capítulo, é apresentado um experimento que envolve os conceitos de multiplicidade preditiva, importância dos atributos para a predição de modelos e métodos de interpretabilidade e, com isso, discute algumas possíveis implicações dos resultados na construção de um ambiente seguro para o uso de Inteligência Artificial, em especial, em domínios críticos como a saúde.

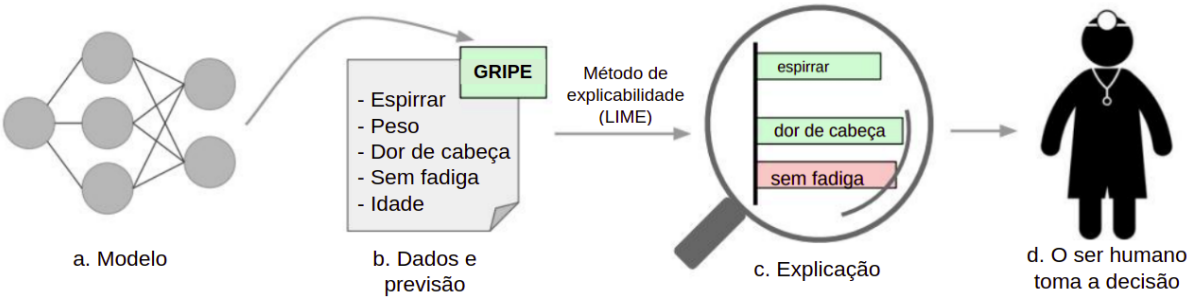
O experimento foi construído por meio da adoção do raciocínio contrafactual, primeiro fixando os resultados dos modelos de predição (classificação) e comparando as listas de importância dos atributos criadas por diferentes métodos de interpretabilidade (contrafactos), segundo fixando os resultados de um modelo específico de predição e métodos de interpretabilidade comparando o resultado das instâncias nos dados (contrafactos).

Para a execução do experimento, foi utilizada a linguagem de programação Python, as bibliotecas scikit-learn[1], PyCaret[2], InterpretML[3] e os métodos de interpretabilidade SHAP[4] e LIME[5].

Sobre interpretabilidade baseada em listas de importância dos atributos, a figura 24.c exibe uma possível explicação para a predição de um modelos. Nela são listados dois atributos que contribuíram positivamente para a predição (espirrar e ter dor de cabeça) e uma em sentido oposto (sem fadiga). Segundo Ribeiro, Singh e Guestrin (2016, p.2), com essa explicação e a predição, um médico pode tomar uma decisão informada sobre confiar na previsão do modelo. Entretanto, métodos de interpretabilidade diferentes podem apresentar listas de importância dos atributos diferentes. Na verdade, um mesmo método de interpretabilidade pode apresentar listas diferentes em função dos hiperparâmetros escolhidos, assim como modelos distintos, mesmo que tenham um desempenho similar, podem conduzir a listas divergentes.

[1] scikit-learn - <https://scikit-learn.org/>
[2] PyCaret - <https://pycaret.gitbook.io/docs/>
[3] InterpretML - <https://interpret.ml/>
[4] SHAP - <https://shap.readthedocs.io/en/latest/index.html>
[5] LIME - <https://github.com/marcotcr/lime>

Figura 24 – Métodos de explicabilidade e importância dos atributos



Fonte: Adaptado pelo autor a partir de Ribeiro, Singh e Guestrin (2016, p.2)

Com o experimento, este capítulo busca evidenciar que uma lista de importância dos atributos pode não ser suficiente para, de forma confiável, apoiar uma decisão informada e, consequentemente, gerar informações que permitam verificar se essa decisão é justa e eticamente aceitável. Apesar dos vários estudos que buscam tornar as explicações mais robustas e estáveis, essa ainda é uma questão que necessita de amadurecimento.

As próximas seções apresentam resultados que mostram como uma lista de importância de atributos pode variar para um mesmo conjunto de dados. No experimento, não são utilizadas técnicas de ataques adversários (*adversarial attacks*) ou variações nos hiperparâmetros dos métodos de interpretabilidade, que poderiam gerar novas listas de importância dos atributos. Deve-se destacar que, por mais que o experimento tenha feito uma comparação entre modelos gerados a partir de algoritmos diferentes, de forma similar, a multiplicidade preditiva poderia ser encontrada em modelos gerados a partir de um único algoritmo, mas configurados com hiperparâmetros diferentes.

5.1 Descrição e planejamento do experimento

O experimento é uma tarefa de classificação e utiliza um conjunto de dados para a predição de diagnóstico de câncer de mama[6]. São trinta atributos preditores e uma variável alvo, com 569 instâncias, sendo 212 (37%) diagnósticos classificados como malignos e 357 (63%) como benignos. Para a fase de pré-processamento, optou-se por um tratamento simples, somente removendo alguns atributos com alta colinearidade e também os que tinham variância muito baixa, além da normalização dos atributos numéricos.

Foram selecionados quatro algoritmos para treinar os modelos de classificação, dois considerados interpretáveis e dois classificados como caixas-pretas. Os interpretáveis são a Regressão Logística[7] (*Logistic Regression* - LR) e *Explainable Boosting Machine*[8] (EBM),

[6] *Breast cancer Wisconsin dataset (classification)* - <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html>

[7] LR - <https://christophm.github.io/interpretable-ml-book/logistic.html>

[8] EBM - Nori et al. (2019)

enquanto os considerados caixas-pretas são o *eXtreme Gradient Boosting*[9] (XGBoost) e o *Random Forest*[10] (RF).

Após a realização das etapas de treino e teste nos 4 (quatro) modelos de classificação, foram aplicados dois métodos para fornecer interpretabilidade de escopo local. O *Local interpretable model-agnostic explanations*[11] (LIME) e o *SHapley Additive exPlanations*[12] (SHAP). O LIME gera modelos substitutos locais transparentes para explicar as previsões individuais. Esses modelos substitutos aproximam o comportamento do modelo caixa-preta em uma determinada região (MOLNAR, 2019). Por outro lado, segundo Molnar (2019), o "objetivo do SHAP é explicar a predição de uma instância x calculando a contribuição de cada atributo para essa predição". Para isso, o método baseia-se na Teoria dos Jogos, tratando cada valor de um atributo da instância x como um participante de uma coalizão e, a partir do cálculo de valores de Shapley (*Shapley values*), é possível determinar a contribuição de cada atributo para a predição.

Em resumo, o planejamento do experimento seguiu as seguintes etapas:

1. Selecionar base de dados;
2. Realizar tratamento de dados, ajustando a escala dos atributos numéricos por meio da normalização e removendo atributos com alta colinearidade ou baixa variância;
3. Dividir os dados em conjuntos de treino (85%) e teste (15%);
4. Treinar os algoritmos selecionados (EBM, LR, RF e XGBoost);
5. Aplicar os métodos de interpretabilidade LIME e SHAP;
6. Avaliação da ocorrência do efeito de multiplicidade preditiva;
7. Seleção de instância de interesse e verificação da consistência nas explicações.

5.2 Resultados do experimento

A tarefa planejada para este experimento é uma classificação binária (diagnóstico benigno ou maligno) e, como o conjunto de dados não está balanceado, optou-se por utilizar a métrica *Area Under the Curve ROC* (AUC) como a mais relevante.

O quadro 2 apresenta as métricas obtidas pelos modelos durante as fases de treinamento e teste, em que todos os modelos obtiveram um excelente desempenho, com diferenças mínimas na métrica AUC.

[9] XGBoost - Chen e Guestrin (2016)
[10] RF - Breiman (2001)
[11] LIME - Ribeiro, Singh e Guestrin (2016)
[12] SHAP - Lundberg e Lee (2017)

Quadro 2 – Performance dos modelos nos dados de treinamento e teste

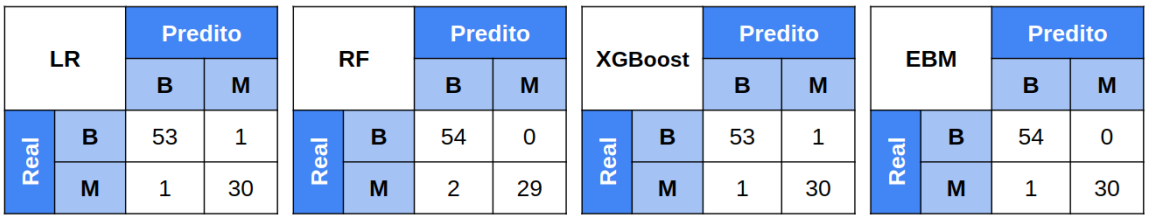
	Treino		Teste	
Modelo	**Acurácia**	**AUC**	**Acurácia**	**AUC**
LR	0,973	0,996	0,977	0,997
RF	0,969	0,995	0,977	0,999
XGBoost	0,961	0,993	0,977	0,998
EBM	0,963	0,991	0,988	0,997

Fonte: Elaborado pelo autor.

5.2.1 Multiplicidade preditiva

A partir da discussão apresentada na seção 4.6, buscou-se verificar a ocorrência da multiplicidade preditiva, ou seja, se modelos concorrentes atribuiriam previsões conflitantes, mas nesse caso, em um cenário com pouca margem para divergências, em função do excepcional desempenho de todos os modelos. A figura 25 exibe as matrizes de confusão para cada um dos modelos com o conjunto de testes. Nele há 85 instâncias, sendo 54 diagnósticos benignos (B) e 31 malignos (M).

Figura 25 – Matrizes de confusão (conjunto de dados de teste)



Fonte: Elaborado pelo autor

No total, com relação ao conjunto de testes, 5 instâncias foram classificadas equivocadamente. O quadro 3 lista os índices das instâncias, com a classe entre parênteses (B ou M), e os relaciona com os modelos responsáveis pelos erros na classificação.

Quadro 3 – Instâncias classificadas com erro

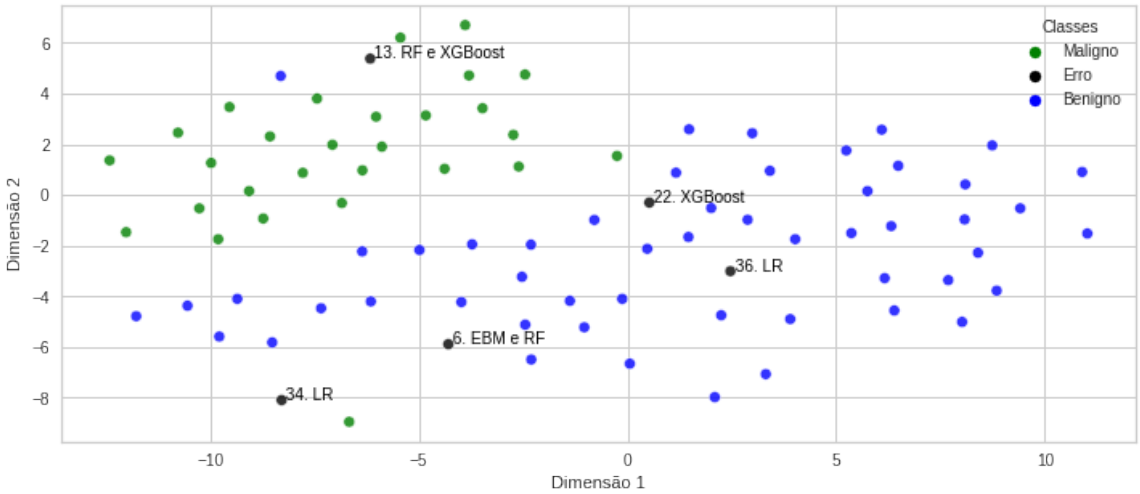
	Instâncias				
Modelos	**6 (M)**	**13 (M)**	**22 (B)**	**34 (M)**	**36 (B)**
LR				X	X
RF	X	X			
XGBoost		X	X		
EBM	X				

Fonte: Elaborado pelo autor.

Na figura 26 são apresentadas as instâncias de teste em uma representação com duas dimensões. Para isso, foi utilizada a técnica de redução de dimensionalidade t-SNE

(*T-distributed Stochastic Neighbor Embedding*) com os parâmetros *default*. O objetivo era criar uma visualização da distribuição dos pontos com erro no conjunto de testes, mesmo que essa técnica de redução de dimensionalidade não permita uma afirmação categórica sobre a existência, ou não, de um padrão na distribuição dos pontos.

Figura 26 – Representação em duas dimensões do conjunto de teste com a técnica t-SNE



Fonte: Elaborado pelo autor

Segundo Marx, Calmon e Ustun (2020, p. 1), a existência de vários modelos igualmente adequados para realizar uma tarefa de *Machine Learning*, mas que atribuem previsões conflitantes, pode levar a desafios éticos. Em casos como esses, o autor afirma que não devemos usar as explicações de um único modelo para tirar conclusões. Para ilustrar essa questão, se o modelo adotado em um serviço de saúde fosse o LR (*logistic regression*) deste experimento, os pacientes representados pelas instâncias 34 e 36 receberiam diagnósticos equivocados (ver quadro 3), diferente do que aconteceria com os outros três modelos (RF, XGBoost e EBM) de desempenho similar. Além do mais, como serviços de saúde podem utilizar modelos diferentes, os mesmos dados um paciente poderiam levar a diagnósticos completamente diferentes em outro serviço de saúde, minando a confiança na tecnologia.

Assim, medir e relatar a multiplicidade preditiva, da mesma forma que é feito com as outras métricas de erros na fase de testes, passa a ser uma questão a ser encarada, pois isso ampliaria as possibilidade de contestação de uma decisão (MARX; CALMON; USTUN, 2020, p. 2). Neste experimento, de um total de 85 instâncias, até cinco poderiam ter o seu diagnóstico afetado (ambiguidade = 5,88%). Além disso, a escolha entre dois modelos diferentes pode alterar o diagnóstico de até quatro instâncias ao mesmo tempo (discrepância = 4,71%). Entretanto, é importante lembrar que os dados utilizados neste experimento permitiram o treino de modelos com desempenho excepcional, o que restringiu a margem para erros. Em casos mais complexos, é provável que as métricas de ambiguidade

e discrepância (ver 4.6) sejam ainda maiores, como em Marx, Calmon e Ustun (2020, p. 7-8) .

5.2.2 Interpretabilidade para um modelo interpretável

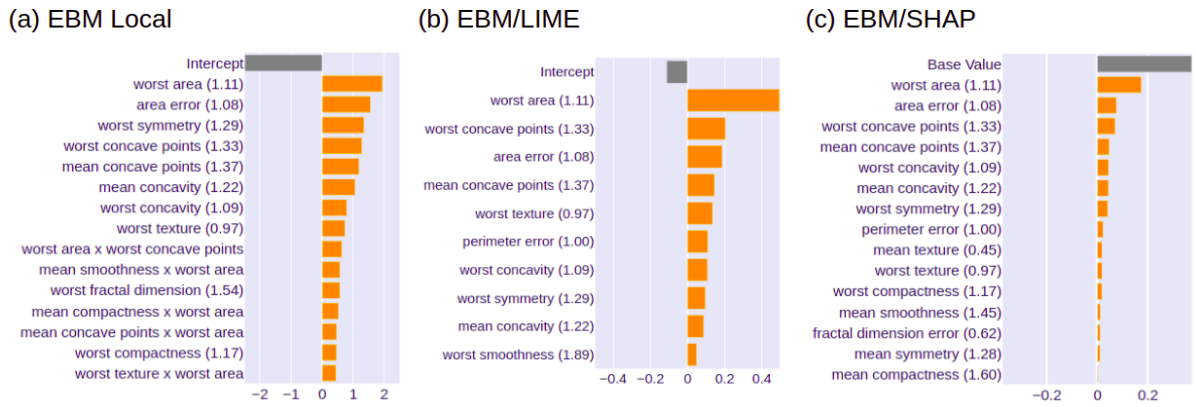
Modelos interpretáveis são capazes de expor a lógica interna que orienta as suas predições, ao contrário dos modelos do tipo caixa-preta. Por outro lado, métodos de interpretabilidade são projetados para identificar os padrões aprendidos por um modelo opaco, mas é razoável esperar que também funcionem com modelos intrinsecamente interpretáveis como árvores de decisão, regressão logística e *Explainable Boosting Machine* (EBM).

Nesta seção, a lista de importância dos atributos fornecida por um modelo intrinsecamente interpretável, treinado com o algoritmo *Explainable Boosting Machine* (EBM), é comparada às listas geradas pelos métodos de interpretabilidade LIME e SHAP, aplicados ao modelo EBM. Desconsiderando os pesos atribuídos, a hipótese utilizada é a de que a ordem de importância dos atributos mais relevantes seja aproximadamente a mesma, se LIME e SHAP puderem interpretar corretamente a estratégia utilizada pelo modelo EBM na classificação das instâncias informadas. Além disso, como o LIME é um método que não gera explicações globais, o objetivo foi restrito a comparar explicações locais para algumas das instâncias do conjunto de testes.

Das 85 instâncias do conjunto de testes, as avaliações se concentraram em um conjunto de 12, que incluem as 5 instâncias classificadas com erros por pelo menos um dos modelos treinados (quadro 3).

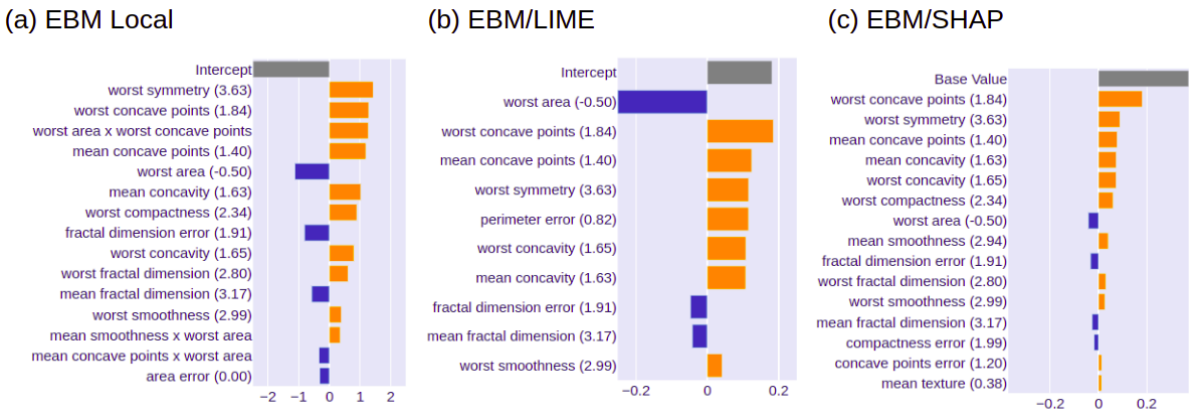
Nas instâncias analisadas, foi possível ver que a divergência era maior a medida que a probabilidade predita se afastava dos extremos (0 e 1). Dois exemplos podem ser vistos nas figuras 27 e 28, com probabilidades preditas de 1,000 e 0,978, respectivamente. Nas figuras são exibidas as listas de importância dos atributos geradas pelo modelo EBM (a), pelo LIME (b) e pelo SHAP (c). As listas exibem os atributos em ordem decrescente de importância.

Figura 27 – Comparação de *feature importance* fornecido pelo modelo EBM e com o uso do LIME e SHAP (instância: 4)



Fonte: Elaborado pelo autor

Figura 28 – Comparação de *feature importance* fornecido pelo modelo EBM e com o uso do LIME e SHAP (instância: 0)



Fonte: Elaborado pelo autor

Com essa análise, a intenção é a de iniciar uma reflexão sobre os limites dos métodos de interpretabilidade para revelar os padrões utilizados pelos modelos, principalmente quando utilizado por profissionais não especializados no tema.

Neste ponto, não é parte integrante dos objetivos deste experimento utilizar essas listas de importâncias dos atributos como elementos para inferência causal, o que quase sempre será inadequado quando feito a partir de um modelo projetado para predição. Entretanto, entender os atributos mais relevantes em uma predição é algo crítico em decisões potencialmente sensíveis e de alto risco, como na saúde, pois limita a capacidade das pessoas afetadas de obter um desfecho favorável e pode atuar como um fator de confusão.

É importante destacar que as explicações fornecidas, mesmo que sejam consistentes, se referem aos padrões identificados pelo modelo treinado, ou seja, não necessariamente as explicações refletirão um mesmo padrão presente nos dados. Como exemplo, um mesmo efeito pode ser capturado de diferentes variáveis com algum nível de correlação e, assim, modelos diferentes pode usar recursos diferentes para uma predição e, mesmo assim, manter um desempenho similar. Em resumo, modelos tentam identificar padrões nos dados, enquanto métodos de interpretabilidade são utilizados para tentar inferir a estratégia de predição dos modelos. Segundo Saarela e Jauhiainen (2021, p. 2), a importância dos atributos dos dados de entrada dependem do modelo de classificação correspondente e uma característica importante para um modelo pode não ser importante para outro modelo.

Nas próximas seções esse aspecto é discutido ao comparar as explicações obtidas com modelos de arquiteturas distintas, sendo alguns do tipo caixa-preta, para os quais se torna bastante complexo compreender de forma plena o seu funcionamento interno.

5.2.3 Importância dos atributos entre modelos e métodos de interpretabilidade diferentes

Importância dos atributos (*feature importance*) é uma das principais estratégias utilizadas pelos métodos de interpretabilidade para fornecer alguma informação sobre os principais atributos utilizados em uma predição (SAARELA; JAUHIAINEN, 2021, p. 1). Entretanto, modelos concorrentes treinados a partir de um mesmo conjunto de dados podem fornecer listas de importância dos atributos diferentes, seja diretamente por serem interpretáveis, seja com o auxílio de métodos de interpretabilidade. Além disso, a própria expressão "importância dos atributos" pode se referir a coisas sutilmente diferentes. Por exemplo, modelos baseados na implementação XGBoost possuem diferentes abordagens para atribuir pesos aos atributos. No quadro 4 é possível ver, em ordem decrescente de relevância, os primeiros 10 atributos para cada um dos principais tipos de atribuição de importância (*gain*, *cover* e *weight*[13]). Considerando somente os 5 primeiros atributos de cada coluna, o único atributo em comum é *worst area*, o que mostra como pode variar o que genericamente se denomina de importância dos atributos.

O quadro 4 mostra três possíveis listas de importância dos atributos que podem ser utilizadas como explicações globais, todas fornecidas pelo próprio modelo. Entretanto, ainda seria possível utilizar um dos vários métodos de interpretabilidade disponíveis para obter novas listas. Assim, o desafio seria escolher e justificar qual delas é a mais adequada. Há diversos trabalhos que se propõem a aumentar a robustez das explicações feitas com importância dos atributos, mas esse é um tema que permanece em debate (ZIEN; KR,

[13] XGBoost *feature importance type* - <https://xgboost.readthedocs.io/en/latest/python/python_api.html?highlight=get_score#xgboost.Booster.get_score>

Quadro 4 – 10 mais importantes atributos para o modelo XGBoost em função do tipo de importância

Importância dos atributos para o modelo XGBoost treinado			
#	***Gain***	***Cover***	***Weight***
1	worst area	worst concave points	worst area
2	mean concave points	fractal dimension error	worst texture
3	worst concave points	smoothness error	area error
4	mean concavity	worst area	worst concave points
5	mean texture	mean concavity	mean texture
6	worst concavity	area error	mean concave points
7	worst texture	mean concave points	worst concavity
8	area error	worst fractal dimension	worst symmetry
9	perimeter error	perimeter error	mean smoothness
10	mean compactness	worst texture	worst smoothness

Fonte: Elaborado pelo autor.

2009; CASALICCHIO; MOLNAR; BISCHL, 2018; FISHER; RUDIN; DOMINICI, 2019; AGARWAL et al., 2022).

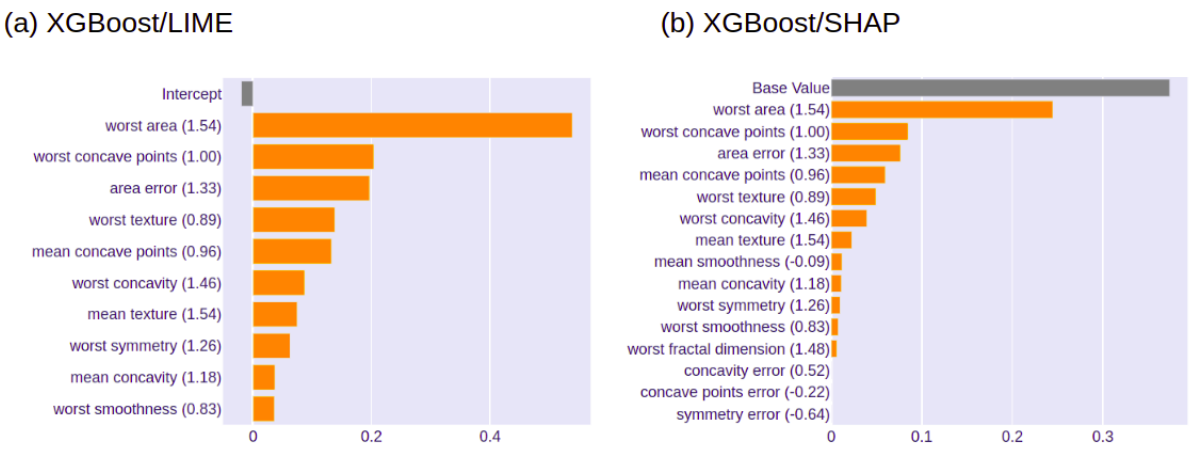
Se por um lado a importância global do atributo mede a importância do recurso para todo o modelo, a importância local refere-se à contribuição de um atributo para os resultados de uma entrada específica. Explicações locais confiáveis são fundamentais para permitir que as partes afetadas possam verificar se estão sendo tratadas de maneira justa, ou, até mesmo, entender como obter um desfecho diferente. Com isso em mente, a próxima etapa do experimento analisa algumas decisões locais e compara as explicações fornecidas por dois dos principais métodos de interpretabilidade (LIME e SHAP).

As figuras 29 e 30 exibem listas de importância dos atributos produzidas pelos métodos LIME (a) e SHAP (b) para o modelo XGBoost treinado neste experimento. A primeira figura exibe a explicação para a instância 5 do conjunto de dados, para a qual o modelo atribuiu uma probabilidade de predição de 1,000. A segunda figura lista os atributos identificados como mais relevantes para a instância 13 (um dos dois erros do XGBoost), com probabilidade de predição de 0,7967.

Dentro do conjunto de 12 instâncias analisadas empiricamente, parece haver um padrão em que, quando o modelo consegue identificar com maior facilidade a classe correta de uma instância, os métodos de interpretabilidade utilizados apresentam listas similares (figura 29). Por outro lado, para as instâncias classificadas incorretamente, LIME e SHAP mantiveram listas similares de importância dos atributos para os modelos EBM e RF, ao contrário do que ocorreu com o XGBoost (figura 30) e LR. Esse comportamento sugere que alguns métodos de interpretabilidade podem ser mais eficientes para identificar o padrão de funcionamento de determinados algoritmos ou que sejam mais adequados ao

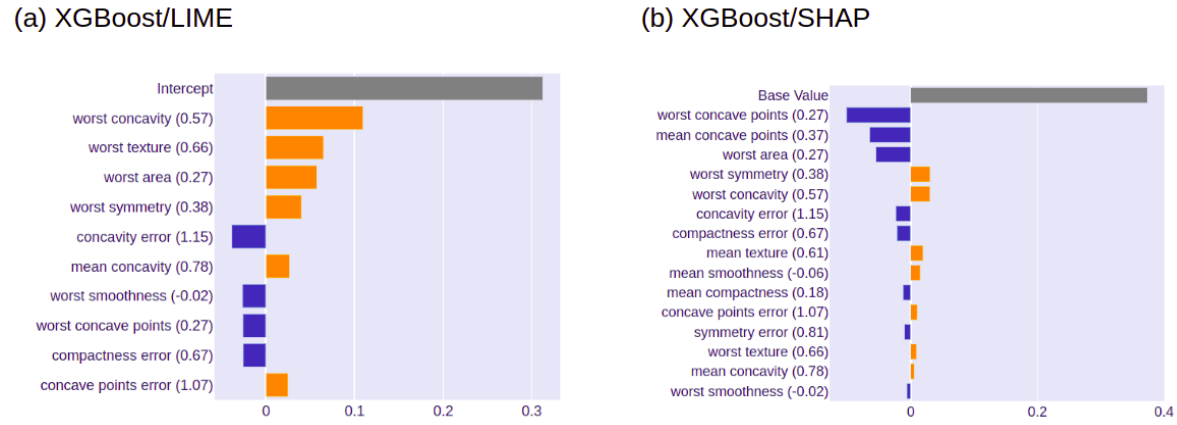
tipo de estratégia que melhor aproxima a função de classificação (linear ou não linear). Entretanto, uma análise mais profunda desta questão não faz parte do escopo desta tese, mas será tratada em trabalhos futuros.

Figura 29 – Comparação de *feature importance* fornecido para o modelo XGBoost com o uso do LIME e SHAP (instância: 5)



Fonte: Elaborado pelo autor

Figura 30 – Comparação de *feature importance* fornecido para o modelo XGBoost com o uso do LIME e SHAP (instância: 13)

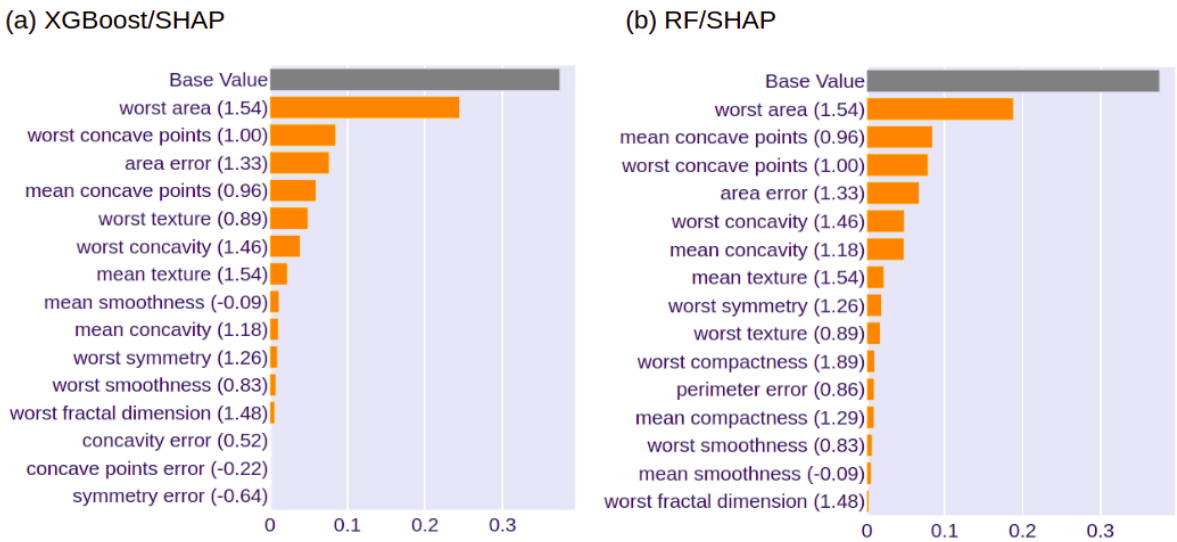


Fonte: Elaborado pelo autor

A análise anterior apresenta o cenário em que dois métodos de interpretabilidade são aplicados a um único modelo. A próxima etapa inverte esse cenário e aplica um método de interpretabilidade a quatro modelos diferentes. A figura 31 exibe as listas de importância dos atributos geradas pelo SHAP para o XGBoost e LR e, de forma similar ao exemplo anterior, as listas de importância dos atributos são semelhantes nas situações em que a incerteza era baixa no momento de classificação. A instância 5 (figura 31) foi predita

corretamente pelos quatro modelos treinados e, em todos os casos, com alta probabilidade (XGBoost, RF e EBM: 1,000; LR: 0,998).

Figura 31 – *Feature importance* gerado com SHAP para XGBoost e RF (instância: 5)

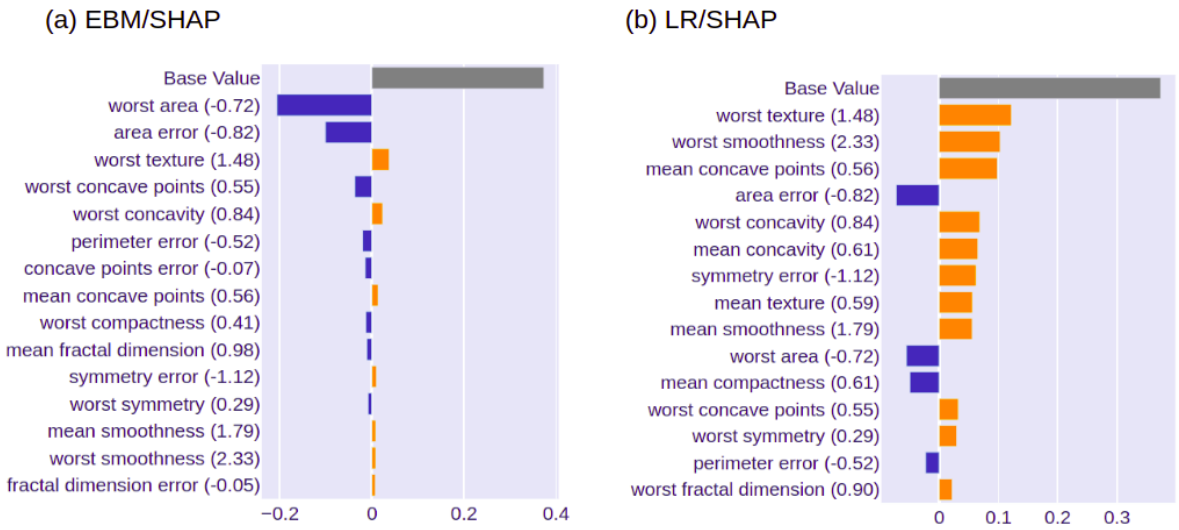


Fonte: Elaborado pelo autor

Em contrapartida, as listas de importância de atributos geradas para a instância 6, classificada com erro pelos modelos EBM e XGBoost, repetem o padrão descrito até aqui, ou seja, quando os modelos erram, ou quando há uma incerteza maior para a classificação de uma instância, as listas de importância dos atributos para os métodos de interpretabilidade são menos coesas.

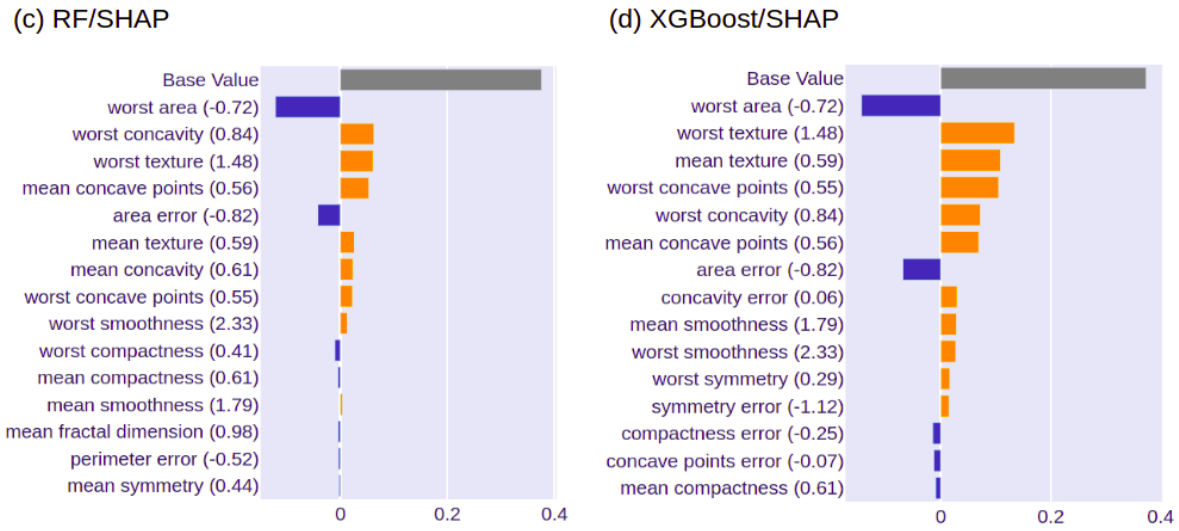
Para ilustrar a afirmação anterior, as figuras 32 e 33 referem-se ao método SHAP aplicado aos quatro modelos treinados. Quanto ao LIME, ele se comportou tão bem quanto o SHAP na instância 5, em que os modelos classificaram corretamente e havia pouca incerteza. Entretanto, na instância 6, o LIME apresentou listas de importância dos atributos mais consistentes do que o SHAP.

Figura 32 – *Feature importance* gerado com SHAP para EBM e LR (instância: 6)



Fonte: Elaborado pelo autor

Figura 33 – *Feature importance* gerado com SHAP para XGBoost e RF (instância: 6)



Fonte: Elaborado pelo autor

Obviamente, nenhuma conclusão contundente pode ser obtida com esta análise empírica, feita a partir de um único conjunto de dados e analisando um grupo limitado de instâncias. No entanto, o objetivo aqui é o de destacar a complexidade de definir e justificar uma lista como a mais adequada para revelar a importância dos atributos para um modelo. Ao mesmo tempo, este experimento ajuda na discussão sobre requisitos para a transparência em serviços que utilizam *Machine Learning*, especialmente quando as explicações podem ser inconsistentes e as pessoas estão sujeitas aos efeitos da multiplicidade preditiva.

5.3 Considerações sobre este capítulo

Aplicações baseadas em *Machine Learning* podem impactar enormemente a sociedade e estão sendo utilizadas em um número cada vez maior atividades distintas. Entretanto, para setores como a saúde e a segurança pública, não basta que o modelo tenha uma boa performance quando a tarefa inclui decisões potencialmente sensíveis ou de alto risco. É necessário compreender quais fatores foram determinantes para uma predição e, com isso, poder avaliar se ela foi justa, se ela não representa um tratamento discriminatório e, entre outras questões, tornar possível a devida prestação de contas e responsabilização (*accountability*).

Como discutido anteriormente, modelos intrinsecamente interpretáveis permitem, com maior facilidade, a compreensão das regras utilizadas em uma predição, enquanto modelos considerados opacos tornam essa uma tarefa muito mais árdua. Para lidar com essa questão, muito esforço tem sido direcionado para o desenvolvimento de métodos de interpretabilidade de modelos opacos (*black box*), o que adiciona mais uma camada de complexidade, que é justificada por alguns autores com a promessa de um melhor desempenho. Outros autores contestam essa afirmação sobre um melhor desempenho e ainda destacam os riscos de usar modelos opacos (RUDIN, 2019; CARUANA, 2019; RUDIN; RADIN, 2019).

O principal artefato gerado por modelos interpretáveis ou por métodos de interpretabilidade é a lista de importância dos atributos (*feature importance*), entretanto, ela pode variar enormemente em função do modelo treinado, dos hiperparâmetros utilizados e do método de interpretabilidade aplicado, ou seja, a questão central é conseguir definir qual lista melhor representa os pesos atribuídos pelo modelo para uma decisão. E nem sempre essa é uma questão evidente para as partes envolvidas ou afetadas por modelos baseados em *Machine Learning*.

Saarela e Jauhiainen (2021, p. 9) fazem um estudo e comparam os seus resultados com o de um experimento anterior. Dentre os 9 atributos identificados como mais relevantes em seu estudo, somente 3 estavam entre os 14 identificados no estudo anterior. Segundo as autoras, isso seria decorrente da existência de atributos altamente correlacionados, o que permite que um resultado possa ser obtido usando diferentes conjuntos de recursos. Entretanto, essa descoberta exige acesso aos dados de treinamento, o que nem sempre é possível, seja por preocupações com a privacidade ou, em especial, pelo uso de soluções proprietárias.

Além disso, ainda neste experimento (ver 5.2.1), foi apresentada a ocorrência da multiplicidade preditiva quando considerados quatro modelos concorrentes (similares e com ótimo desempenho). O objetivo neste ponto foi o de chamar atenção para a necessidade de discussão sobre o tratamento que deve ser dado a essa questão. No campo da saúde, é

razoável esperar que situações idênticas tenham o mesmo tratamento, mas isso pode não acontecer quando um indivíduo é afetado pelos efeitos da multiplicidade preditiva.

Este capítulo mostrou que a lista de importância dos atributos pode variar de inúmeras formas e, em situações de alto impacto, isso pode ser uma barreira para a construção de um ambiente confiável e seguro para a IA na saúde. É importante destacar que a questão pode ser ainda mais complexa quando é colocada neste contexto a possibilidade de manipulação de explicações Lakkaraju e Bastani (2020), Dombrowski et al. (2019), Slack et al. (2021).

6 Discussão

Nesta tese, a partir das hipóteses de pesquisa (ver seção 1.4.2), são estabelecidas as linhas de discussões principais sobre o uso de IA na saúde. A primeira se relaciona com um olhar para um quadro mais amplo, em que modelos baseados em IA/ML podem ser vistos como apenas um dos componentes ou artefatos de uma solução tecnológica proposta (ver capítulos 2 e 3). Nesta tese, afirma-se que garantir um tratamento justo e não discriminatório não é uma tarefa limitada a questões técnicas, mas envolve compreender os fatores externos que condicionam a solução, além dos elementos que podem contribuir para mitigar riscos. Nesse sentido, a regulação, a interdisciplinaridade e a alfabetização (*literacy*) sobre a IA assumem um papel de destaque. Na próxima seção (6.1), foram incluídos alguns novos aspectos à discussão já apresentada sobre o tema.

A segunda questão move o foco para esse componente baseado em IA, que em muitos casos é considerado pouco transparente, seja pela arquitetura escolhida, pela complexidade envolvida ou por aspectos ligados à proteção da propriedade intelectual. Entretanto, em campos complexos e de alto impacto como a saúde, é importante que seja possível identificar os fatores que influenciaram a decisão, inclusive para assegurar que o tratamento é justo e não discriminatório. Assim, dada a opacidade de alguns modelos, a transparência assume um papel central e, com isso, os métodos de interpretabilidade tornam-se elementos fundamentais para a construção de confiança e para viabilizar a responsabilização (*accountability*) de decisões algorítmicas (capítulo 4). Infelizmente, em muitas situações, é uma tarefa complexa assegurar que as explicações apresentadas são suficientemente robustas.

Para mitigar alguns efeitos do uso de modelos do tipo caixa-preta, evidências recentes mostram que em muitos casos os modelos intrinsecamente interpretáveis conseguem desempenho similar aos melhores modelos opacos. Assim, esta tese se alinha à ideia de que os modelos interpretáveis devem ser priorizados, especialmente em situações de grande impacto, como é comum na saúde.

Por fim, no capítulo 5, o experimento apresentado utiliza a discussão feita nos capítulos anteriores e destaca alguns pontos de fragilidade quando a opção é pelo uso de modelos opacos (*black box*), mas sem perder de vista que a simples troca por modelos interpretáveis não garante decisões justas e não discriminatórias, o que ainda dependerá de elementos externos à IA, como a regulação e, até mesmo, a definição de justiça adotada.

Neste capítulo, a seção 6.2 retoma a discussão sobre os métodos de interpretabilidade, transparência e modelos intrinsecamente interpretáveis e a seção 6.3, a partir desta pesquisa, apresenta propostas para a discussão sobre o uso da IA na saúde, de forma que, orientado

pelo princípio da Equidade do SUS, a IA possa contribuir para a redução das iniquidades sociais e econômicas da sociedade, não para perpetuá-las.

6.1 A construção de um ambiente confiável e justo para o uso de ML na saúde não está restrito às questões técnicas

Há uma enorme expectativa de que a Inteligência Artificial (IA), ou mais especificamente *Machine Learning* (ML) e *Deep Learning* (DL), possam contribuir cada vez mais para a melhoria dos serviços de saúde, reduzindo custos e ampliando o acesso. No entanto, fazer o melhor uso dessa tecnologia exige uma busca contínua por melhores e mais robustas soluções tecnológicas, mas elas dificilmente serão suficientes para garantir que essas soluções tratem de forma equânime, justa e não discriminatória as pessoas e grupos afetados. Isso depende de definições e elementos externos à IA, que moldam a sua construção e aplicação.

Reconhecer a influência desses elementos externos pode ser fundamental para a construção de um ambiente mais justo, pois isso pode facilitar a atribuição de responsabilidades entre os diversos componentes de uma solução. Por exemplo, não é a IA que define os objetivos do projeto, a população alvo, os dados que serão coletados, como serão tratados ou como as predições serão utilizadas. Entretanto, talvez em função da popularidade obtida por aplicações exitosas nos mais diversos campos, utilizar algum dispositivo ou serviço baseado em IA parece atrair muita atenção e, em certo casos, críticas. Infelizmente, como mostra o caso de prevenção de gravidez precoce (capítulo 2), uma solução adequada depende também de muitos fatores externos, além das questões técnicas. Logo, remover o uso de IA não necessariamente tornará a solução mais justa, embora ela possa contribuir para tornar menos transparente os processos que conduziram a uma decisão.

Nesse contexto, a discussão sobre as situações em que o uso de IA é admissível deve envolver, sempre que possível, uma comparação tomando como linha de base a tarefa executada por humanos.

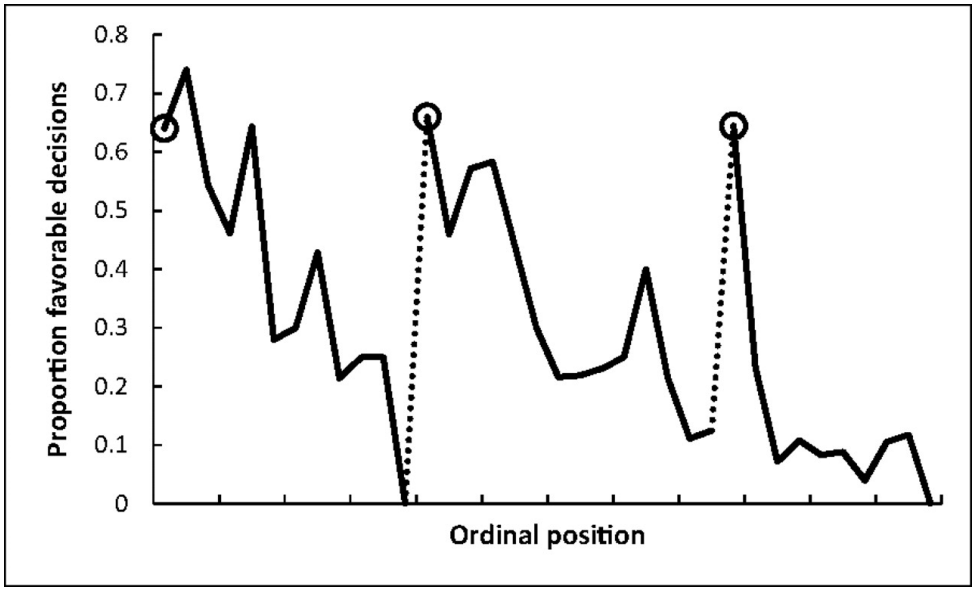
6.1.1 Decisões humanas e modelos preditivos

Imaginemos, por hipótese, que exista um modelo de ML que, a partir dos dados fornecidos, decide se concede ou não um pedido de liberdade condicional para um condenado a pena privativa de liberdade. Em geral, 35% dos pedidos são concedidos. No entanto, descobriu-se que, além dos dados do condenado, o número de decisões tomadas desde a última pausa do sistema influencia enormemente a predição. Assim, os casos julgados no início de um ciclo possuem uma probabilidade maior de um desfecho favorável, ao contrário daquele julgados próximos de uma pausa na operação do sistema. Seria razoável

manter esse modelo de ML em produção?

Este exemplo é hipotético, mas baseado em um estudo com seres humanos apresentado por Danziger, Levav e Avnaim-Pesso (2011). Nele, são analisadas as decisões de juízes experientes no julgamento de pedidos de liberdade condicional. Os pesquisadores relatam que as decisões favoráveis aos réus atingem picos de aproximadamente 65% no início do dia e após as pausas, chegando a quase zero no final de cada ciclo. A figura 34 ilustra essa questão exibindo as proporções de decisões favoráveis aos detentos por posição ordinal. Os pontos circulados indicam a primeira decisão em cada uma das três sessões. As marcas na escala no eixo x denotam cada terceiro caso e a linha pontilhada indica pausa para alimentação.

Figura 34 – Concessão de liberdade condicional ao longo do dia



Fonte: Danziger, Levav e Avnaim-Pesso (2011, p. 2)

Discutindo esse estudo, Kahneman (2012, p. 50) destaca que "juízes cansados e com fome tendem a incorrer na mais fácil posição *default* de negar os pedidos de condicional". Danziger, Levav e Avnaim-Pesso (2011) sugerem que este comportamento para as decisões esteja relacionado ao esgotamento dos juízes em função das seguidas decisões tomadas. Os autores acreditam que um efeito similar deva ocorrer com outros especialistas, como nas decisões médicas ou financeiras.

Utilizando o exemplo sobre concessão de liberdade condicional, é possível selecionar alguns elementos para o debate de como comparar decisões humanas com as tomadas por modelos de ML. Neste sentido, quanto mais próximo das condições reais de uso, mais cedo e, de forma mais segura, pode-se implantar ou descartar o uso de ML em decisões sensíveis e de alto impacto. Destaca-se que, descartar o uso de ML por ele não possuir uma performance muito superior a um grupo de especialistas, pode representar a perda da

uma oportunidade de ampliar o acesso a um determinado serviço, especialmente se esses especialistas não estão disponíveis para toda a população.

Por outro lado, não é razoável comparar decisões humanas com as de um modelo de ML em condições de laboratório, pois o funcionamento pode ser muito diferente em condições reais. Por exemplo, ao apresentar um estudo com um modelo de detecção de retinopatia diabética, Beede et al. (2020, p. 6) relata a ocorrência de limitações para o seu funcionamento adequado por falta de iluminação correta na unidade de saúde, o que pode alterar completamente a precisão do preditor, quando comparada com a obtida em laboratório. Quanto aos humanos, comparações que não levem em consideração fatores como sobrecarga, exaustão e vieses cognitivos, talvez não sejam realistas.

Segundo Simon (1991), March (1978), os indivíduos pretendem ser racionais (*intended rationality*), porém sofrem restrições relativas à sua "capacidade cognitiva e informacional". O conceito de racionalidade limitada está relacionado, principalmente, às seguintes restrições (que podem ocorrer conjunta ou separadamente): (a) problemas de atenção: tempo e capacidade de atenção são limitados. Nem tudo pode ser resolvido ao mesmo tempo; (b) problemas de memória: as capacidades dos indivíduos para armazenar informações são limitadas; (c) problemas de compreensão: os decisores têm dificuldade em organizar, resumir e utilizar informações para formar inferências sobre as conexões causais de eventos e sobre as características dos problemas; e (d) problemas de comunicação: capacidade limitada para comunicar e compartilhar informações técnicas e complexas (Simon,1991, March,1978 apud Pedroso, 2014, p. 63).

Obviamente, saber quando um modelo é melhor que seres humanos para uma tarefa não é simples. Análises feitas a partir de dados e decisões do COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) apontaram a existência de um forte viés racial[1] contra negros. Dressel e Farid (2018), ao avaliar se esse modelo era melhor do que humanos para estimar a probabilidade de reincidência dos condenados, o que justificaria, em parte, o uso de uma solução que torna o processo menos transparente, afirmou que o software utilizado não era melhor do que a avaliação de pessoas com pouca ou nenhuma experiência em justiça criminal. Entretanto, dois anos depois, Lin et al. (2020) afirmam em outro estudo que, em condições reais, um algoritmo teria uma performance superior aos seres humanos. Em resumo, não é simples a definição de quando um algoritmo pode apoiar ou substituir humanos em determinada tarefa, mas essa discussão deve ser colocada em um espaço transparente, com todas as partes envolvidas, que reconheça as imperfeições de ambos os lados, assim como as vantagens.

Neste ponto, sob a perspectiva de que a IA no SUS deve olhar para desafios em um país de dimensões continentais e com enormes desigualdades, essa discussão tenta chamar a atenção para a necessidade de construção de metodologias mais próximas do mundo real,

[1] ProPublica - <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

para avaliar as situações que podem ser interessantes e seguras para o uso da IA. Afinal, há situações para as quais a IA pode não ser adequada, e forçar o seu uso pode trazer riscos difíceis de serem identificados. Ao mesmo tempo, esperar por condições perfeitas e ideais para a adoção de aplicações de IA pode impedir a expansão do acesso a alguns serviços de saúde.

6.1.2 Regulação

Com o uso de Inteligência Artificial (IA) nos mais diversos campos, há uma crescente preocupação sobre como tornar segura a adoção da tecnologia, especialmente quando ela é aplicada em contextos sensíveis como a saúde, segurança pública e educação (OBERMEYER et al., 2019; MORLEY et al., 2020; FIRTH-BUTTERFIELD et al., 2022). Para lidar com essa questão, inúmeras iniciativas de governos, empresas privadas e organizações da sociedade civil buscam estabelecer diretrizes e princípios para guiar o desenvolvimento de uma IA ética. Entretanto, esse esforço parece não ter sido suficiente para lidar com esse desafios (HAGENDORFF, 2020, p. 10). Segundo Mittelstadt (2019, p. 1):

> As iniciativas existentes para codificar a ética da IA não estão isentas de críticas. Muitas iniciativas, particularmente aquelas patrocinadas pela indústria, têm sido caracterizadas como mera sinalização de virtude destinada a retardar a regulamentação e focar preventivamente o debate em problemas abstratos e soluções técnicas. As iniciativas de ética em IA produziram, até agora, princípios vagos e de alto nível e declarações de valor que prometem orientar a ação, mas na prática fornecem poucas recomendações específicas e não abordam tensões normativas e políticas fundamentais incorporadas em conceitos-chave (por exemplo, justiça, privacidade)[2].

Neste ambiente, avança a discussão em diversas partes do mundo sobre como regular o desenvolvimento da IA, garantindo um ambiente seguro para a sua aplicação. Na seção 3.5.2 foram apresentados algumas dessas abordagens e lá é possível ver que o Brasil ainda se encontra em um estágio inicial desse processo. Infelizmente, nesta pesquisa, não foi possível identificar nenhum esforço coordenado e amplo para a construção de um arcabouço regulatório para a adoção da IA pelo Sistema Único de Saúde (SUS). Acreditamos que o SUS deve ter um papel ativo para garantir uma adoção segura de Inteligência Artificial, identificando os desafios que podem ser específicos para o setor saúde.

[2] *Existing initiatives to codify AI Ethics are not without their critics. Many initiatives, particularly those sponsored by industry, have been characterised as mere virtue signalling intended to delay regulation and pre-emptively focus debate on abstract problems and technical solutions. AI Ethics initiatives have thus far largely produced vague, high-level principles and value statements which promise to be action- guiding, but in practice provide few specific recommendations5 and fail to address fundamental normative and political tensions embedded in key concepts (e.g. fairness, privacy)*

Como exemplo, é importante que seja definido, no ambiente da saúde, como será a atribuição de responsabilidade em decisões que envolvam IA. No caso de algum dano ser causado a um paciente, quem deve ser responsabilizado? O médico? A empresa responsável pela solução tecnológica? O serviço de saúde? Essa não é uma questão completamente clara em muitos outros setores, por isso é importante que se discuta internamente que requisitos e evidências devem estar presentes em função dos possíveis impactos de uma decisão em serviços de saúde.

A responsabilização passa pela definição do quão influente é o papel da IA em uma decisão. Legislações como a *General Data Protection Regulation*[3] (GDPR - Europa) e a Lei Geral de Proteção de Dados Pessoais[4] (LGPD - Brasil) diferenciam as decisões totalmente automatizadas das tomadas com apoio da IA. Segundo Wachter, Mittelstadt e Russell (2017), a GDPR só prevê o direito à explicação em decisões totalmente automatizadas e que tenham efeito legal ou outro efeito similar. De forma semelhante, o artigo 20 da LGPD só prevê a solicitação de revisão de "decisões tomadas unicamente com base em tratamento automatizado de dados pessoais". Esta é uma discussão que deveria ser melhor aprofundada no campo da saúde, pois é razoável esperar que a influência dos dispositivos de IA aumente à medida que eles se tornem mais confiáveis, o que pode se transformar em uma barreira ao direito de revisão de decisões.

Em resumo, sem um papel ativo na construção da regulação do uso da IA, restará ao SUS utilizar as regras e restrições construídas para outros setores, que podem não ser adequadas para a dimensão e as restrições de um setor tão sensível para a sociedade como o da saúde.

As próximas duas seções (6.1.3 e 6.1.4) discutem dois outros tópicos que poderiam, a partir de princípios do SUS como a Equidade, fomentar o diálogo e contribuir para um esforço de regulação da IA.

6.1.3 O que motiva a escolha por um modelo baseado em IA?

Quais devem ser os requisitos para que um modelo de *Machine Learning* seja adotado em um serviço de saúde? A discussão em 4.7 sobre o estudo de Caruana et al. (2015) destaca uma situação em que os pesquisadores optaram por um modelo de menor performance, mas transparente (regressão logística), no lugar de um modelo opaco (rede neural). Com uso de um outro modelo interpretável, os pesquisadores encontraram uma regra que associava baixa probabilidade de óbito a pessoas com pneumonia e histórico de asma. Na verdade, pessoas nessa situação, eram internadas diretamente na UTI e esse tratamento tornava a probabilidade de óbito dessas pessoas menor do que a população em geral.

[3] GDPR - <https://gdpr-info.eu/>
[4] LGPD - <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>

O perigo de usar um modelo opaco, que aprendeu esta regra sem que saibamos, seria o de não internar pacientes que teriam uma probabilidade muito maior de ir a óbito. A questão que se coloca aqui é saber, na falta de regulação, a quem cabe decidir que modelo é aceitável? A quem cabe monitorar o uso desses modelos?

Para mitigar alguns desses riscos, vários autores argumentam que para decisões sensíveis, deve-se usar modelos interpretáveis, evitando modelos opacos que necessitam de métodos de interpretabilidade para fornecer alguma informação sobre o seu funcionamento interno (RUDIN; RADIN, 2019; RUDIN, 2019). Infelizmente, esses métodos de interpretabilidade podem não ser consistentes com os padrões aprendidos pelos modelos e, além disso, alvo de manipulação (LAKKARAJU; BASTANI, 2020; DOMBROWSKI et al., 2019).

Diversos autores afirmam que é possível obter, em muitas situações, um desempenho similar aos melhores modelos caixa-preta com modelos interpretáveis (NORI et al., 2019; RUDIN, 2019). Entretanto, a técnica estar disponível não assegura o seu uso, por isso essa é uma questão que pode ser discutida e tratada por meio da regulação do setor.

6.1.4 Implantação e monitoramento de IA

Uma característica importante de modelos baseados em IA é o fato de que eles tendem a mudar o seu comportamento ao longo do tempo, adaptando-se aos novos dados fornecidos. Essa é uma vantagem dessa tecnologia, mas ela, ao mesmo tempo, cria uma preocupação que vai além do momento em que é avaliada a sua adoção. Modelos com essa característica devem contar com algum tipo de monitoramento, para assegurar que continuem em conformidade com os requisitos estabelecidos no momento da implantação.

Outra questão relevante é a possibilidade de mudança de conceito (*concept drift*). Por exemplo, um modelo que prevê a probabilidade de óbito de pacientes por COVID-19 pode, ao longo do tempo, ver enormes alterações na qualidade das suas predições por eventos externos ao modelo, como o surgimento de uma nova variante do vírus ou o desenvolvimento de uma novo medicamento. Nestes casos, o modelo pode ser ajustado para o novo contexto, mas será necessária uma nova avaliação para verificar se ele continua adequado para a tarefa.

Enfim, garantir a segurança de uma aplicação com IA não se restringe ao momento de implantação, é uma tarefa contínua e esses custos devem ser levados em conta ao avaliar a sustentabilidade da solução tecnológica.

6.1.5 Interdisciplinaridade e letramento em IA

A Ciência de Dados é um campo de estudos em que a interdisciplinaridade ocupa um lugar de destaque. Este tema, em conjunto com a necessidade de letramento (*literacy*),

foi discutido na seção 3.6 e deve se tornar cada vez mais relevante à medida que cresce o uso da IA nos mais diversos setores da sociedade.

Assim, a interdisciplinaridade assume um papel fundamental quando o olhar para uma solução tecnológica baseada em IA se desloca para um quadro mais amplo (ver seção 3.7), em que a IA passa a ser vista como um dos componentes desta solução, mas que pode ser condicionada por decisões externas ao contexto técnico, como a definição da população alvo, a escolha dos dados que serão utilizados na criação do modelo e o uso que será feito das predições deste modelo.

É importante destacar que a diversidade de saberes (interdisciplinaridade) pode contribuir muito para uma IA mais ética, mas a diversidade pode ser pensada de uma forma mais ampla, com relação à origem, raça, gênero, idade, etc. Segundo Firth-Butterfield et al. (2022), a democratização do acesso à IA, e o seu uso seguro e responsável, passam por promover letramento universal em IA e pela priorização da diversidade no seu desenvolvimento e implantação.

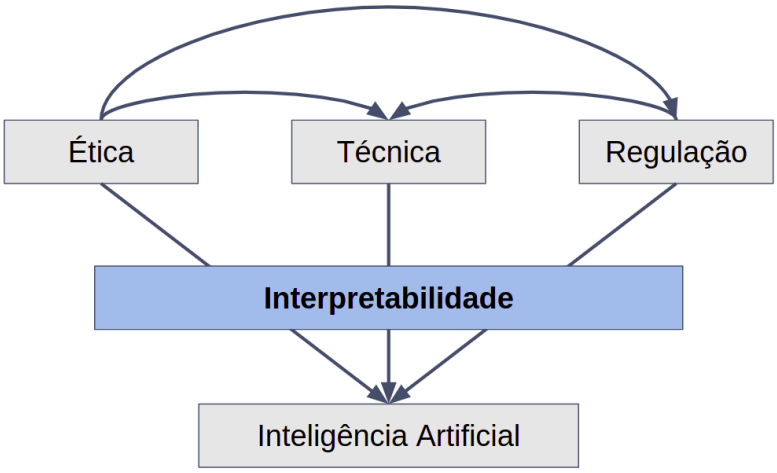
Com relação ao letramento em IA, um estudo apresentado por Prado, Münch e Villarroel (2022), sobre a formação de juízes brasileiros para o uso ético da IA no Judiciário, apontou que "61,3% dos juízes relataram se sentir totalmente despreparados para exercer o controle e a supervisão de sistemas auxiliares de IA para a elaboração de decisões judiciais". Logo, parece razoável supor que o quadro entre os profissionais de saúde não seja muito diferente, o que pode ser uma barreira para a adoção da IA na saúde, especialmente se a expectativa for a de um uso crítico da ferramenta, o que é fundamental em um ambiente com decisões tão sensíveis como às relacionadas à saúde.

Garantir um uso seguro e ético da IA na sociedade depende de muitos fatores, mas certamente passa pela capacitação das pessoas que irão interagir ou que serão afetadas por ela.

6.2 Um ambiente justo e confiável para ML na saúde depende de processos e artefatos robustos que assegurem a transparência

A construção de um ambiente justo e confiável para decisões de alto impacto depende de artefatos que executem de forma transparente as suas tarefas, em que seja possível uma compreensão razoável dos motivos que conduziram a uma decisão. Neste sentido, a interpretabilidade de decisões algorítmicas torna-se fundamental para a verificar conformidade com os objetivos e valores pactuados.

Figura 35 – O papel da interpretabilidade na interação entre IA e aspectos éticos, técnicos e regulatórios



Fonte: Elaborado pelo autor

Como exemplo, a interpretabilidade é fundamental para verificar se a IA atende aos requisitos estabelecidos pela regulação, pois as métricas de acurácia podem não ser suficientes. É importante notar que pode haver uma interação entre uma demanda que tem origem em uma questão ética, que pode gerar a necessidade de regulação e, com isso, exigir intervenções no domínio técnico (ver figura 35). No entanto, sem interpretabilidade consistente e confiável, talvez não seja possível assegurar que as alterações foram eficazes.

Além disso, segundo Carvalho, Pereira e Cardoso (2019, p. 10-11), a interpretabilidade traz, dentre outros, o benefício de promover a aceitação social, a segurança dos modelos, a detecção de comportamentos falhos e, para a pesquisa científica, a interpretabilidade pode contribuir para extração do conhecimento capturado pelos modelos. Além disso, para Rudin, Wang e Coker (2020, p. 6), "não importa qual definição técnica de justiça seja escolhida, é mais fácil debater a justiça de um modelo transparente do que um modelo proprietário" e, certamente, o mesmo se aplica a modelos do tipo caixa-preta, mesmo que não sejam proprietários, pois eles necessitam de uma camada extra de complexidade para fornecer alguma transparência, com o uso de métodos de interpretabilidade (como, por exemplo, os já mencionados LIME e SHAP).

No entanto, Kaur et al. (2020, p. 1) apresentam um estudo em que afirmam que os "resultados indicam que os cientistas de dados confiam demais e fazem uso indevido de ferramentas de interpretabilidade", o que pode comprometer a confiabilidade nas explicações apresentadas. Infelizmente, como discutido nos capítulos 4 e 5, não é simples obter explicações consistentes e confiáveis.

Por isso, esta tese se alinha aos autores que defendem que a opção seja, sempre que possível, pelo uso de modelos intrinsecamente interpretáveis, que podem, em muitos

casos, ter desempenho similar aos modelos opacos (RUDIN, 2019; NORI et al., 2019).

6.2.1 Métodos de interpretabilidade e manipulação de explicações

Um questão importante sobre o uso de métodos de interpretabilidade em conjunto com modelos caixa-preta, no lugar da opção por modelos interpretáveis, é que as explicações derivadas de métodos de interpretabilidade podem estar sujeitas à manipulação por ataques adversários (*adversarial attacks*), o que pode minar a confiança na solução tecnológica (SLACK et al., 2020; FINLAYSON et al., 2019; SLACK et al., 2021).

Além disso, ataques adversários podem possibilitar a prática de *fairwashing*, em que um fornecedor pode manipular as explicações para promover a falsa percepção de que um modelo de ML respeita alguns valores éticos (AÏVODJI et al., 2019). Aïvodji et al. (2021, p. 1) afirma que "esse ataque pode afetar significativamente os indivíduos que receberam um resultado negativo seguindo a previsão do modelo, privando-os da possibilidade de contestá-lo".

Para Anders et al. (2020, p. 1), os métodos de explicação prometem tornar os classificadores caixa-preta mais transparentes e, com isso, eles possam servir "de prova para um processo sensível, justo e confiável de tomada de decisão do algoritmo e, assim, aumentar sua aceitação pelos usuários finais". Segundo o autor, "essas esperanças são atualmente infundadas" (ANDERS et al., 2020, p. 1) e ele afirma em sua conclusão que "métodos de explicação amplamente utilizados não devem ser usados como prova de um processo de tomada de decisão algorítmico justo e sensato. Isso porque eles podem ser facilmente manipulados como demonstramos tanto teórica quanto experimentalmente" (ANDERS et al., 2020, p. 9).

Muitas pesquisas têm sido feitas no campo de estudos da *Explainable Artificial Intelligence* (XAI) com métodos de interpretabilidade, especialmente para tentar tornar esses métodos mais robustos. Entretanto, os autores citados nessa seção alertam para os riscos de manipulação das explicações e algumas de suas consequências. Combinado com isso, não se deve perder de vista os resultados apresentados por Kaur et al. (2020), de que "cientistas de dados confiam demais e fazem uso indevido de ferramentas de interpretabilidade".

Em resumo, um ambiente confiável para o uso da IA na saúde passa necessariamente por explicações também confiáveis. O que se torna mais complexo quando a opção é por modelos caixa-preta e, para obter algum entendimento sobre como as decisões são tomadas, são utilizados métodos de interpretabilidade que, em teoria, podem ser manipulados.

6.2.2 Multiplicidade preditiva e justiça algorítmica

Quando é proposta a discussão de como pode ser construído um ambiente justo e não discriminatório para o uso de IA na saúde, os conceitos de multiplicidade preditiva (ver 4.6) chamam a atenção para o fato de que modelos diferentes, mesmo que tenham um desempenho bom e similar, podem classificar de forma diversa uma mesma pessoa representada nos dados. Esta discussão se articula com o necessário debate sobre quais critérios devem ser utilizados para a seleção de um modelo para uso em ambientes com decisões sensíveis (ver 6.1.3).

Marx, Calmon e Ustun (2020) apresentam duas métricas para medir o impacto da multiplicidade preditiva: ambiguidade e discrepância. Segundo os autores:

> [...] uma ambiguidade de 44% significa que se poderia produzir explicações conflitantes para 44% das previsões. Embora cada explicação nos ajudaria a entender como os modelos concorrentes operam, a evidência de previsões conflitantes forneceria uma salvaguarda contra racionalizações injustificadas. (MARX; CALMON; USTUN, 2020, p. 7-8, tradução nossa)

Quando combinados os efeitos da multiplicidade preditiva com a possível falta de consistência entre métodos de interpretabilidade (ver capítulo 5), o resultado pode ser a criação de um ambiente não confiável para aplicação da IA na saúde, especialmente para as pessoas afetadas por ela.

Por isso, dentre outras propostas já apresentadas ao longo deste capítulo, acreditamos que avaliar a multiplicidade preditiva, antes de selecionar um modelo para a entrada em produção, é um importante caminho para mitigar riscos e construir um ambiente mais justo e responsável para o uso de IA no SUS.

6.3 Propostas

Nesta seção, de forma resumida, são reunidas algumas propostas apresentadas ao longo desta tese com o objetivo de contribuir para a construção de um ambiente mais justo e responsável para a aplicação de IA no Sistema Único de Saúde:

1. Planejamento transparente dos investimentos em recursos humanos e tecnológicos para a adoção de IA no SUS;
2. Em decisões sensíveis, a prioridade deve ser dada para o uso de modelos interpretáveis;
3. A existência de multiplicidade preditiva deve ser avaliada e as medidas para mitigar os seus efeitos devem ser transparentes;
4. Esforços devem ser feitos no sentido de garantir, o mais cedo possível, um olhar interdisciplinar e diverso para projetos que envolvam adoção de IA;

5. Uma estratégia nacional de letramento (*literacy*) deve ser implementada com foco na operação, mas também em um uso crítico da solução tecnológica;

6. Construção de um cadastro de aplicações que utilizem IA em desenvolvimento ou em operação no SUS, tornando possível rastrear e comparar os impactos positivos e riscos envolvidos;

7. Definição de um processo para avaliação, validação e monitoramento de tecnologias baseadas em IA no SUS com um olhar interdisciplinar;

8. Avaliação da Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (Conitec[5]) para adoção de IA no SUS por meio da emissão de relatório técnico sobre a tecnologia avaliada, levando em consideração as evidências científicas, a interpretabilidade, a avaliação econômica e o impacto da incorporação da tecnologia no Sistema Único de Saúde.

Na busca pela construção de um ambiente responsável para o desenvolvimento da IA no SUS, acreditamos que essas propostas formam um conjunto de temas importantes para a discussão, mas certamente muitos outros aspectos podem ser acrescentados para englobar questões não envolvidas nesta pesquisa. Inclusive, parte das preocupações que tivemos sobre os desafios éticos, técnicos ou regulatórios podem não estar explicitamente nessas propostas, mas estão presentes nas discussões apresentadas, especialmente nos capítulos 2 e 3.

Por fim, esperamos que esta tese possa contribuir para a discussão sobre a adoção da Inteligência Artificial no Sistema Único de Saúde, objetivando a busca por soluções tecnológicas justas e não discriminatórias e, com isso, possa atuar na direção da redução das desigualdades, e não para perpetuá-las.

[5] A Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (Conitec) foi criada pela Lei nº 12.401, de 28 de abril de 2011, que dispõe sobre a assistência terapêutica e a incorporação de tecnologia em saúde no âmbito do Sistema Único de Saúde.

A Conitec tem por objetivo assessorar o Ministério da Saúde (MS) nas atribuições relativas à incorporação, exclusão ou alteração de tecnologias em saúde pelo SUS, bem como na constituição ou alteração de protocolo clínico ou de diretriz terapêutica. Disponível em: <http://conitec.gov.br/entenda-a-conitec-2>

7 Conclusão

Esta pesquisa busca contribuir para a discussão sobre o uso de dispositivos e serviços baseados em Inteligência Artificial (IA) em campos potencialmente sensíveis e de alto risco como a saúde, com foco especial no Sistema Único de Saúde (SUS) e em um de seus princípios fundamentais ou doutrinários: a Equidade. Embora a Equidade não conste explicitamente na Lei Orgânica da Saúde[1], as preocupações que se relacionam a ela encontravam-se pressentes no pensamento sanitário brasileiro. E, depois de 1992, ela começa a aparecer nos relatórios das conferências nacionais de saúde (BARROS; SOUSA, 2016, p. 3). Segundo Corrêa Matta (2007):

> O princípio da equidade é fruto de um dos maiores e históricos problemas da nação: as iniquidades sociais e econômicas. Essas iniquidades levam a desigualdades no acesso, na gestão e na produção de serviços de saúde. Portanto, o princípio da equidade, para alguns autores, não implica a noção de igualdade, mas diz respeito a tratar desigualmente o desigual, atentar para as necessidades coletivas e individuais, procurando investir onde a iniquidade é maior.

Assim, esta tese procura compreender como as oportunidades e riscos derivados da adoção da IA podem impactar na sociedade, positivamente ou negativamente, especialmente nas populações mais vulneráveis. Embora nem todos os casos estudados ou as questões aqui colocadas estejam diretamente relacionadas ao domínio da saúde, o SUS e o princípio da Equidade orientaram toda a sua construção.

A pesquisa começa na busca por entender quais requisitos tornariam a IA mais segura e confiável para ser aplicada no SUS, mas havia a hipótese de que fatores externos (não técnicos) poderiam influenciar decisivamente a construção e o uso da solução. Neste ponto do estudo, surge a necessidade de um olhar para um quadro mais amplo e, assim, a pesquisa se expande para entender o papel da interdisciplinaridade, da regulação (capítulo 3) e dos fatores que orientam a construção do componente baseado em IA. Focar nesse componente não explicaria as decisões de quais dados coletar, qual parcela da população seria afetada e como seriam utilizados os resultados do modelo que será treinado e aplicado.

O estudo de predição de gravidez precoce (capítulo 2) destaca muitos desses pontos, especialmente a forma como uma visão sexista da questão da gravidez não atribuiu um papel para os meninos na predição (2.3.2), recaindo toda a vigilância sobre as meninas. Ou seja, a discussão sobre ética e justiça não pode ficar restrita à questão técnica, é necessário um olhar mais amplo. Neste cenário, um programa como o de predição de gravidez precoce

[1] Lei 8080/90 - <http://www.planalto.gov.br/ccivil_03/leis/L8080.htm>

deve ser visto como um conjunto de componentes e a IA é somente um deles, incapaz de garantir que todo o programa seja adequado, quando observado o princípio da Equidade.

Depois desse olhar para o quadro mais amplo, a pesquisa se concentra em um elemento fundamental para validar modelos treinados e aplicados em decisões sensíveis ou de alto risco, garantindo que eles ajam da forma esperada: a interpretabilidade das decisões algorítmicas, foco do capítulo 4. Muitos modelos utilizados atualmente são pouco transparentes e, quando apresentam explicações, nem sempre elas são estáveis e confiáveis. Além disso, não costuma haver uma grande transparência sobre as decisões tomadas no treinamento do modelo e que conduziram àquela explicação, especialmente quando se trata de uma solução proprietária.

Mesmo quando não há limitação de acesso aos dados e algoritmos, os métodos de interpretabilidade podem gerar explicações divergentes (capítulo 5), o que pode se tornar uma barreira ao desenvolvimento de confiança na solução e, em especial, torna frágil qualquer afirmação sobre o quão justo é o tratamento de um modelo, independente da noção de justiça que seja adotada. Assim, do ponto de vista do princípio da Equidade, decisões sem explicações consistentes não permitem aferir se a solução está contribuindo para diminuir ou para perpetuar iniquidades sociais.

Por fim, esta pesquisa reconhece as enormes oportunidades que o uso da IA pode representar no campo da saúde, ao mesmo tempo em que destaca alguns riscos importantes que precisam ser mitigados, para que a IA não se torne uma ferramenta para a perpetuação de iniquidades sociais e que possa contribuir para o desenvolvimento do Sistema Único de Saúde e de toda a sociedade.

7.1 Discussão sobre as hipóteses da pesquisa

Hipótese 1: *O desenvolvimento de um ambiente confiável e justo para a aplicação de Machine Learning na Saúde inclui as questões tecnológicas, mas não está restrito a elas.*

O estudo realizado nesta tese corroborou a hipótese 1 da pesquisa, ao identificar diversos fatores externos que podem influenciar IA e, com isso, demonstrou que a discussão sobre a construção de um ambiente mais justo e seguro para o uso de IA depende de muitos fatores não técnicos, como apontado nos capítulos 2 e 3 e na seção 6.1.

Hipótese 2: *Métodos de interpretabilidade distintos podem gerar explicações substancialmente diferentes, mesmo quando aplicados a casos idênticos.*

Dada a complexidade envolvida nas aplicações de IA, o papel da interpretabilidade torna-se central para o uso da tecnologia. Inclusive, para promover a aceitação social, a

segurança dos modelos, a detecção de comportamentos falhos e, para a pesquisa científica, a interpretabilidade pode contribuir para o estabelecimento de hipóteses a partir de dados. Nesta tese, o estudo realizado corroborou a hipótese 2 da pesquisa, ao discutir os principais métodos de interpretabilidade e verificar a existência de fenômenos como a multiplicidade preditiva e inconsistências entre métodos interpretabilidade, quando comparadas as explicações sobre as decisões tomadas. As inconsistências encontradas podem ocultar tratamentos discriminatórios ou, por falta de confiança na solução, funcionar como uma barreira para a adoção da tecnologia, o que pode, por exemplo, levar a perda de enormes oportunidades para a melhoria do acesso aos serviços de saúde. A discussão sobre a segunda hipótese de pesquisa encontra-se, principalmente, nos capítulos 4 e 5 e na seção 6.2.

7.2 Respostas para as questões de pesquisa

Questão 1: *Ao buscar a construção de um ambiente o mais confiável e justo possível para a aplicação de* Machine Learning *na Saúde, quais aspectos devem ser observados e quais atores devem estar presentes?*

Para responder a essa questão de pesquisa, esta tese parte analisa os fatores que interagem para a criação da solução tecnológica com o uso de IA. A hipótese inicial era a de que empregar os melhores e mais sofisticados recursos tecnológicos não forneceria necessariamente uma solução tecnológica justa e segura para uso no campo da saúde.

Diversos fatores externos interagem para moldar esse ambiente, como a regulação, o uso de saberes interdisciplinares, a existência de diversidade nas equipes ou o letramento (*literacy*) das pessoas envolvidas. Outro fator importante é a visão política que orienta o processo, pois ela pode estabelecer a população alvo do projeto, os dados que serão utilizados e como as predições da IA serão utilizadas (ver seção 2.5).

Desta forma, entendendo a IA como apenas um componente de um quadro mais amplo, é possível ver as limitações da tecnologia para garantir que a solução final seja justa e atenda aos padrões éticos esperados. Entretanto, além do fato de que a construção de um ambiente confiável e seguro para o uso da IA não deva estar limitada às questões técnicas, o envolvimento de todas as partes interessadas, especialmente das pessoas afetadas pelas decisões, é fundamental para estabelecer a confiança no processo e para mitigar riscos.

Questão 2: *As estratégias selecionadas, propostas para a interpretabilidade de modelos de* Machine Learning *são adequadas para aplicações no campo da Saúde?*

O uso de IA na Saúde pode promover enormes avanços, mas garantir que o seu uso seja justo e não discriminatório contra pessoas, grupos, comunidades, populações e

instituições, depende da transparência de como são tomadas as decisões. A opacidade inerente a algumas aplicações da IA impede a compreensão do funcionamento interno dos modelos e, consequentemente, pode tornar pouco consistente o processo de avaliação da qualidade da solução.

Para lidar com essa questão, nos últimos anos foram desenvolvidos diversos métodos para fornecer alguma interpretabilidade para modelos opacos (*black boxes*). No entanto, esta abordagem depende da consistência e da confiança que pode ser depositada nas explicações fornecidas. Por outro lado, diversos autores argumentam que, para decisões sensíveis e de alto risco, devem ser priorizados algoritmos que intrinsecamente transparentes.

A partir da pesquisa e do experimento realizado para a elaboração desta tese, não foi possível afirmar que as explicações apresentadas, com o uso de dois dos principais métodos de interpretabilidade, sejam consistentes com os padrões aprendidos pelos modelos de IA. Assim, a interpretabilidade obtida pode não ser suficiente para a atribuição de responsabilidades por essas decisões (*accountability*) ou a avaliação de sua conformidade a padrões e normas legais, técnicas e éticas. Além disso, explicações não consistentes podem ser utlizadas para a prática de *fairwashing*, em que um fornecedor pode manipular as explicações para promover a falsa percepção de que um modelo de ML respeita alguns valores éticos. Por isso, esta tese se alinha com os autores que defendem o uso prioritário de modelos intrinsecamente transparentes, especialmente em decisões sensíveis ou de alto risco.

Referências

ABRÀMOFF, M. D.; TOBEY, D.; CHAR, D. S. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *American Journal of Ophthalmology*, p. 1–9, 2020. ISSN 18791891. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32171769/>. Citado 2 vezes nas páginas 34 e 35.

AGARWAL, C. et al. Rethinking Stability for Attribution-based Explanations. p. 1–9, 2022. Disponível em: <http://arxiv.org/abs/2203.06877>. Citado 2 vezes nas páginas 75 e 88.

AGARWAL, R. et al. Neural Additive Models: Interpretable Machine Learning with Neural Nets. n. NeurIPS, 2020. Disponível em: <http://arxiv.org/abs/2004.13912>. Citado na página 75.

AÏVODJI, U. et al. Fairwashing: The risk of rationalization. *36th International Conference on Machine Learning, ICML 2019*, v. 2019-June, p. 240–252, 2019. Citado 2 vezes nas páginas 72 e 103.

AÏVODJI, U. et al. Characterizing the risk of fairwashing. n. NeurIPS, 2021. Disponível em: <http://arxiv.org/abs/2106.07504>. Citado 2 vezes nas páginas 72 e 103.

ALVAREZ-MELIS, D.; JAAKKOLA, T. S. On the Robustness of Interpretability Methods. n. Whi, 2018. Disponível em: <http://arxiv.org/abs/1806.08049>. Citado na página 73.

ANDERS, C. J. et al. Fairwashing explanations with off-manifold detergent. *37th International Conference on Machine Learning, ICML 2020*, PartF16814, p. 291–300, 2020. Citado 2 vezes nas páginas 73 e 103.

BABIC, B. B. et al. Beware explanations from AI in health care the benefits of explainable artificial intelligence are not what they appear. *Science*, v. 373, n. 6552, p. 284–286, 2021. ISSN 10959203. Disponível em: <https://www.science.org/doi/10.1126/science.abg1834>. Citado 3 vezes nas páginas 70, 71 e 72.

BAKER, A. et al. A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. *Frontiers in Artificial Intelligence*, v. 3, p. 9, 2020. ISSN 26248212. Disponível em: <https://www.frontiersin.org/articles/10.3389/frai.2020.543405/full>. Citado na página 35.

Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, Elsevier B.V., v. 58, n. December 2019, p. 82–115, 2020. ISSN 15662535. Disponível em: <https://doi.org/10.1016/j.inffus.2019.12.012>. Citado 3 vezes nas páginas 13, 43 e 62.

BARROS, F. P. C. de; SOUSA, M. F. de. Equidade: Seus conceitos, significações e implicações para o SUS. *Saude e Sociedade*, v. 25, n. 1, p. 9–18, 2016. ISSN 01041290. Disponível em: <https://doi.org/10.1590/S0104-12902016146195>. Citado na página 106.

BEEDE, E. et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: *CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, US: [s.n.], 2020. p. 12. ISBN 9781450367080. Disponível em: <https://doi.org/10.1145/3313831.3376718>. Citado 3 vezes nas páginas 35, 36 e 97.

BOCCOLINI, C. S. et al. Fatores associados à discriminação percebida nos serviços de saúde do Brasil: Resultados da Pesquisa Nacional de Saúde, 2013. *Ciencia e Saude Coletiva*, v. 21, n. 2, p. 71–78, 2016. ISSN 16784561. Disponível em: <https://doi.org/10.1590/1413-81232015212.19412015>. Citado na página 15.

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. October 2001, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 82.

BROWN, L. et al. Challenging the Use of Algorithm-driven Decision-making in Benefits Determinations Affecting People with Disabilities. *Center for Democracy and Technology*, 2020. Disponível em: <https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-p>. Citado 2 vezes nas páginas 39 e 40.

BUDD, J. et al. Digital technologies in the public-health response to COVID-19. *Nature Medicine*, Springer US, v. 26, n. 8, p. 1183–1192, 2020. ISSN 1546170X. Disponível em: <http://dx.doi.org/10.1038/s41591-020-1011-4>. Citado na página 35.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: PMLR. *Conference on fairness, accountability and transparency*. 2018. p. 77–91. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. Citado na página 41.

CALMON, F. P. et al. Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, v. 2017-Decem, n. Nips, p. 3993–4002, 2017. ISSN 10495258. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf>. Citado na página 63.

CARUANA, R. Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning. In: *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. New York, NY, USA: Association for Computing Machinery, 2019. Disponível em: <https://dl.acm.org/doi/10.1145/3292500.3340414>. Citado 2 vezes nas páginas 44 e 92.

CARUANA, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, v. 2015-Augus, p. 1721–1730, 2015. Disponível em: <https://dl.acm.org/doi/10.1145/2783258.2788613>. Citado 2 vezes nas páginas 77 e 99.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, v. 8, n. 8, 2019. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/8/8/832>. Citado 3 vezes nas páginas 63, 66 e 102.

CASALICCHIO, G.; MOLNAR, C.; BISCHL, B. Visualizing the feature importance for black box models. In: *Lecture Notes in Computer Science*. Springer International Publishing, 2018. v. 11051 LNAI, p. 655–670. ISBN 9783030109240. ISSN 16113349. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-10925-7_40>. Citado na página 88.

CHEN, C. et al. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, v. 32, n. NeurIPS, p. 1–12, 2019. ISSN 10495258. Disponível em: <https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf>. Citado na página 77.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, v. 13-17-August-2016, p. 785–794, 2016. Citado na página 82.

CNJ. *Poder Judiciário Conselho Nacional de Justiça RESOLUÇÃO N*. Brasília: Conselho Nacional de Justiça, 2020. 1–11 p. Disponível em: <https://atos.cnj.jus.br/atos/detalhar/3429>. Citado 2 vezes nas páginas 50 e 51.

COLLERA, V. *Nossos engenheiros no Google não estudaram filosofia, temos que ajudá-los*. 2019. Disponível em: <https://brasil.elpais.com/brasil/2019/01/28/eps/1548684447_982945.html>. Citado na página 55.

Corrêa Matta, G. Princípios e Diretrizes do Sistema Único de Saúde. *Políticas de saúde: organização e operacionalização do Sistema Único de Saúde*, p. 61–80, 2007. Disponível em: <https://www.arca.fiocruz.br/handle/icict/39223>. Citado na página 106.

DANZIGER, S.; LEVAV, J.; AVNAIM-PESSO, L. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, v. 108, n. 17, p. 6889–6892, 2011. ISSN 00278424. Disponível em: <https://doi.org/10.1073/pnas.1018033108>. Citado na página 96.

DASTIN, J. *Amazon scraps secret AI recruiting tool that showed bias against women | Reuters*. 2018. Disponível em: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>. Citado na página 41.

DAVANCENS, F. *Predicción de Embarazo Adolescente con Machine Learning*. 2017. Disponível em: <https://github.com/facundod/case-studies/blob/master/Prediccionde EmbarazoAdolescenteconMachineLearning.md>. Citado na página 24.

De Cerqueira, J. A. S.; TIVES, H. A.; CANEDO, E. D. Ethical Guidelines and Principles in the Context of Artificial Intelligence. *ACM International Conference Proceeding Series*, n. i, 2021. Citado na página 46.

DECARIO, N.; ETZIONI, O. *America Needs AI Literacy Now — pnw.ai*. 2021. Disponível em: <https://pnw.ai/article/america-needs-ai-literacy-now/72515409>. Citado 2 vezes nas páginas 56 e 57.

DOMBROWSKI, A. K. et al. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, v. 32, p. 1–34, 2019. ISSN 10495258. Citado 3 vezes nas páginas 72, 93 e 100.

DOMBROWSKI, A. K. et al. Towards robust explanations for deep neural networks. *Pattern Recognition*, Elsevier Ltd, v. 121, p. 108194, 2022. ISSN 00313203. Disponível em: <https://doi.org/10.1016/j.patcog.2021.108194>. Citado na página 72.

DOSHI-VELEZ, F.; KIM, B. Towards A Rigorous Science of Interpretable Machine Learning. n. Ml, p. 1–13, feb 2017. Disponível em: <http://arxiv.org/abs/1702.08608>. Citado 3 vezes nas páginas 61, 62 e 64.

DRESSEL, J.; FARID, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, v. 4, n. 1, p. 1–6, 2018. ISSN 23752548. Citado na página 97.

ELEBI, C. M. Inteligencia Artificial y Salud. p. 58, 2020. Disponível em: <http://hdl.handle.net/10908/17691>. Citado 2 vezes nas páginas 28 e 35.

European Commission. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. *Com(2021)*, v. 0106, p. 1–108, 2021. Citado na página 53.

FINLAYSON, S. G. et al. Emerging vulnerabilities demand new conversations. *Science*, v. 363, n. 6433, p. 1287–1290, 2019. Disponível em: <http://science.sciencemag.org/content/363/6433/1287>. Citado na página 103.

FIRTH-BUTTERFIELD, K. et al. *Without universal AI literacy, AI will fail us | World Economic Forum*. 2022. Disponível em: <https://www.weforum.org/agenda/2022/03/without-universal-ai-literacy-ai-will-fail-us/>. Citado 5 vezes nas páginas 40, 49, 56, 98 e 101.

FISHER, A.; RUDIN, C.; DOMINICI, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, v. 20, p. 1–81, 2019. ISSN 15337928. Citado 3 vezes nas páginas 70, 73 e 88.

FJELD, J. et al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet & Society*, p. 110, 2020. Disponível em: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420>. Citado 2 vezes nas páginas 46 e 47.

G1-PB. *Campina Grande testa projeto pioneiro para monitoramento tecnológico do 'Criança Feliz'*. 2019. Disponível em: <https://g1.globo.com/pb/paraiba/noticia/2019/09/25/campina-grande-testa-projeto-pioneiro-para-monitoramento-tecnologico-do-crianca-feliz.ghtml>. Citado na página 29.

GOH, K. H. et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, v. 12, n. 1, p. 1–10, 2021. ISSN 20411723. Disponível em: <https://www.nature.com/articles/s41467-021-20910-4.pdf>. Citado 2 vezes nas páginas 37 e 38.

GOMES, A. L. *Parceria entre governo brasileiro, província argentina e Microsoft — Português (Brasil)*. 2009. Disponível em: <https://www.gov.br/cidadania/pt-br/noticias-e-conteudos/desenvolvimento-social/noticias-desenvolvimento-social/parceria-entre-governo-brasileiro-provincia-argentina-e-microsoft>. Citado na página 29.

Gonçalves Escarião, P. H. et al. Epidemiologia e diferenças regionais da retinopatia diabética em Pernambuco, Brasil. *Arquivos Brasileiros de Oftalmologia*, v. 71, n. 2, p. 172–175, 2008. ISSN 00042749. Disponível em: <https://doi.org/10.1590/S0004-27492008000200008>. Citado na página 35.

GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM Computing Surveys*, v. 51, n. 5, 2018. ISSN 15577341. Disponível em: <https://dl.acm.org/doi/10.1145/3236009>. Citado na página 33.

HABLI, I.; LAWTON, T.; PORTER, Z. Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization*, World Health Organization, v. 98, n. 4, p. 251–256, apr 2020. ISSN 15640604. Citado na página 71.

HAGENDORFF, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, Springer Netherlands, v. 30, n. 1, p. 99–120, 2020. ISSN 15728641. Disponível em: <https://doi.org/10.1007/s11023-020-09517-8>. Citado 3 vezes nas páginas 48, 49 e 98.

High Level Expert Group on AI. *Assessment List for Trustworthy AI (ALTAI)*. [s.n.], 2020. 0–33 p. ISBN 9789276200093. Disponível em: <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>. Citado na página 53.

JEMIO, D.; HAGERTY, A.; ARANDA, F. *The Case of the Creepy Algorithm That 'Predicted' Teen Pregnancy | WIRED*. 2022. Disponível em: <https://www.wired.com/story/argentina-algorithms-pregnancy-prediction/>. Citado na página 20.

KAHNEMAN, D. *Rápido e Devagar: Duas formas de pensar*. Rio de Janeiro: Objetiva, 2012. 588 p. ISSN 1098-6596. ISBN 9788539004010. Citado na página 96.

KALIL, A. J. et al. Sepsis risk assessment: A retrospective analysis after a cognitive risk management robot (Robot Laura®) implementation in a clinical-surgical unit. *Research on Biomedical Engineering*, v. 34, n. 4, p. 310–316, 2018. ISSN 24464740. Disponível em: <https://doi.org/10.1590/2446-4740.180021>. Citado na página 36.

KAUR, H. et al. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Conference on Human Factors in Computing Systems - Proceedings*, p. 1–14, 2020. Citado 3 vezes nas páginas 67, 102 e 103.

LAKKARAJU, H.; BASTANI, O. "how do i fool you?": Manipulating user trust via misleading black box explanations. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, p. 79–85, 2020. Citado 2 vezes nas páginas 93 e 100.

LECHER, C. *A healthcare algorithm started cutting care, and no one knew why - The Verge*. 2018. Disponível em: <https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy>. Citado 2 vezes nas páginas 39 e 40.

LEDFORD, H. Millions of black people affected by racial bias in health-care algorithms. *Nature*, NLM (Medline), v. 574, n. 7780, p. 608–609, oct 2019. ISSN 14764687. Disponível em: <https://www.nature.com/articles/d41586-019-03228-6>. Citado na página 42.

LIAA. *Sobre la predicción automática de embarazos adolescentes*. Laboratorio de Inteligencia Artificial Aplicada (LIAA), 2018. Disponível em: <https://liaa.dc.uba.ar/es/sobre-la-prediccion-automatica-de-embarazos-adolescentes/>. Citado 4 vezes nas páginas 21, 24, 25 e 26.

LIN, Z. J. et al. The limits of human predictions of recidivism. *Science Advances*, v. 6, n. 7, p. 1–9, 2020. ISSN 23752548. Citado na página 97.

LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, Association for Computing Machinery, New York, NY, USA, v. 16, n. 3, p. 31–57, jun 2018. ISSN 1542-7730. Disponível em: <https://doi.org/10.1145/3236386.3241340>. Citado 4 vezes nas páginas 42, 44, 61 e 62.

LIU, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license, v. 1, n. 6, p. e271–e297, 2019. ISSN 25897500. Disponível em: <http://dx.doi.org/10.1016/S2589-7500(19)30123-2>. Citado na página 34.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964. Disponível em: <https://dl.acm.org/doi/10.5555/3295222.3295230>. Citado na página 82.

MARCH, J. G. Bounded rationality, ambiguity, and the engineering of choice. *The bell journal of economics*, JSTOR, p. 587–608, 1978. Disponível em: <http://ipwna.ir/wp-content/uploads/2018/04/Bounded_Rationality_Ambiguity-irpublicpolicy.pdf>. Citado na página 97.

MARX, C. T.; CALMON, F. D. P.; USTUN, B. Predictive multiplicity in classification. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-9, p. 6721–6730, 2020. Disponível em: <https://proceedings.mlr.press/v119/marx20a.html>. Citado 5 vezes nas páginas 73, 74, 84, 85 e 104.

MCTI. *Estratégia Brasileira de Inteligência Artificial - EBIA* . 2021. Disponível em: <https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/inteligencia-artificial>. Citado na página 50.

MITTELSTADT, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, v. 1, n. 11, p. 501–507, 2019. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293>. Citado 3 vezes nas páginas 48, 58 e 98.

MOLNAR, C. *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. [s.n.], 2019. Disponível em: <https://christophm.github.io/interpretable-ml-book>. Citado 3 vezes nas páginas 66, 68 e 82.

MORLEY, J. et al. The ethics of AI in health care: A mapping review. *Social Science and Medicine*, v. 260, n. June, 2020. ISSN 18735347. Citado 3 vezes nas páginas 41, 59 e 98.

National Health Expenditure. *NHE Fact Sheet | CMS*. 2021. Disponível em: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet>. Citado 2 vezes nas páginas 17 e 34.

News Center Microsoft Latinoamérica. *Avanza el uso de la Inteligencia Artificial en la Argentina con experiencias en el sector público, privado y ONGs*. 2018. Disponível em: <https://news.microsoft.com/es-xl/avanza-el-uso-de-la-inteligencia-artificial-en-la-argentina-con-experiencias-en-el-sector-publico-privado>. Citado 2 vezes nas páginas 21 e 22.

NORI, H. et al. InterpretML: A Unified Framework for Machine Learning Interpretability. 2019. Disponível em: <http://arxiv.org/abs/1909.09223>. Citado 4 vezes nas páginas 76, 81, 100 e 103.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, p. 447–453, 2019. ISSN 10959203. Citado 2 vezes nas páginas 42 e 98.

Ortiz Freuler, J.; IGLESIAS, C. *Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay*. [S.l.], 2018. 36 p. Disponível em: <http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf>. Citado 11 vezes nas páginas 21, 22, 23, 24, 25, 26, 28, 121, 122, 123 e 124.

Parliament UK. *House of Lords - AI in the UK: ready, willing and able? - Artificial Intelligence Committee*. 2017. Disponível em: <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>. Citado na página 61.

PCDAS. *Ciência de Dados aplicada à Saúde*. 2021. Disponível em: <https://pcdas.icict.fiocruz.br/ciencia-de-dados-aplicada-a-saude/>. Citado 2 vezes nas páginas 54 e 55.

PEDROSO, M. Racionalidade limitada e uso de informações técnicas em modelos de análise de políticas públicas: proposições sobre a perspectiva integradora da Análise Multicritério de Decisão Espacial Construtivista. *RP3 - Revista de Pesquisa em Políticas Públicas*, v. 0, n. 2, p. 59–83, 2014. Disponível em: <https://periodicos.unb.br/index.php/rp3/article/view/14600/12911>. Citado na página 97.

PEDROSO, M. D. E. M. *Inteligência Decisória E Análise De Políticas Públicas*. Tese (Tese de Doutorado) — Universidade de Brasília, 2011. Disponível em: <https://repositorio.unb.br/handle/10482/9663>. Citado na página 33.

PEÑA, P.; VARON, J. *Gravidez na adolescência abordada pelo colonialismo de dados de um sistema que é patriarcal desde o projeto*. 2021. Disponível em: <https://notmy.ai/pt/noticias/gravidez-na-adolescencia-abordada-pelo-colonialismo-de-dados-de-um-sistema-que-e-patriarcal-desde->. Citado 4 vezes nas páginas 21, 24, 29 e 30.

PERRAULT, R. et al. Artificial Intelligence Index 2019 Annual Report. *Standford University - Human-Centered Artificial Intelligence*, p. 291, 2019. Disponível em: <https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf>. Citado 2 vezes nas páginas 47 e 48.

PRADO, E. M. B.; MÜNCH, L. A. C.; VILLARROEL, M. A. C. U. "Sob controle do usuário": formação dos juízes brasileiros para o uso ético da IA no Judiciário. *Revista do Tribunal Regional Federal da 4ª Região*, Porto Alegre, RS, 2022. Disponível em: <https://www.trf4.jus.br/trf4/controlador.php?acao=pagina_visualizar&id_pagina=2287>. Citado 2 vezes nas páginas 51 e 101.

PUREAI. *Researchers Release Open Source Counterfactual Machine Learning Library*. 2020. Disponível em: <https://pureai.com/articles/2020/03/13/open-source-counterfactuals.aspx>. Citado na página 68.

QAYYUM, A. et al. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE REVIEWS IN BIOMEDICAL ENGINEERING*, v. 14, p. 156–180, 2021. Disponível em: <https://tinyurl.com/FDA-AI-diabetic-eye>. Citado na página 35.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should i trust you?"Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, v. 13-17-Augu, p. 1135–1144, 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939778>. Citado 4 vezes nas páginas 65, 80, 81 e 82.

RODRIGUES, N. et al. Análise de tendência de mortalidade por sepse no Brasil e por regiões de 2010 a 2019. *Revista de Saúde Pública*, v. 56, p. 1–13, 2022. Disponível em: <https://doi.org/10.11606/s1518-8787.2022056003789>. Citado na página 36.

RONG, Y. et al. Evaluating Feature Attribution: An Information-Theoretic Perspective. 2022. Disponível em: <https://arxiv.org/abs/2202.00449>. Citado na página 75.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, v. 1, n. 5, p. 206–215, 2019. ISSN 25225839. Disponível em: <https://arxiv.org/pdf/1811.10154.pdf>. Citado 8 vezes nas páginas 18, 45, 67, 76, 77, 92, 100 e 103.

RUDIN, C.; RADIN, J. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, v. 1, n. 2, 11 2019. Https://hdsr.mitpress.mit.edu/pub/f9kuryi8. Disponível em: <https://hdsr.mitpress.mit.edu/pub/f9kuryi8>. Citado 3 vezes nas páginas 14, 92 e 100.

RUDIN, C.; WANG, C.; COKER, B. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, v. 2, n. 1, p. 1–60, 2020. Disponível em: <https://hdsr.mitpress.mit.edu/pub/7z10o269>. Citado 2 vezes nas páginas 78 e 102.

Ruth Hailu. *Fitbits, other wearables may not accurately track heart rates in people of color*. 2019. Disponível em: <https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/>. Citado na página 41.

RYAN, M.; STAHL, B. C. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, v. 19, n. 1, p. 61–86, 2021. ISSN 17588871. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html>. Citado na página 50.

SAARELA, M.; JAUHIAINEN, S. Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, Springer International Publishing, v. 3, n. 2, p. 1–12, 2021. ISSN 25233971. Disponível em: <https://doi.org/10.1007/s42452-021-04148-9>. Citado 3 vezes nas páginas 69, 87 e 92.

SANDLER, R.; BASL, J. *BUILDING DATA AND AI ETHICS COMMITTEES*. 2019. 26 p. Disponível em: <https://www.accenture.com/_acnmedia/PDF-107/Accenture-AI-And-Data-Ethics-Committee-Report-11.pdf>. Citado na página 55.

SIMON, H. A. Bounded rationality and organizational learning. *Organization science*, INFORMS, v. 2, n. 1, p. 125–134, 1991. Disponível em: <https://pubsonline.informs.org/doi/10.1287/orsc.2.1.125>. Citado na página 97.

SINGH, S. *Next Wave of Artificial Intelligence Market Worth $190B by 2025*. 2019. Disponível em: <https://www.prnewswire.com/news-releases/next-wave-of-artificial-intelligence-market-worth-190b-by-2025--exclusive-study-by-marketsandmarket
html>. Citado na página 56.

SLACK, D. et al. Fooling LIME and SHAP. p. 180–186, 2020. Citado na página 103.

SLACK, D. et al. *Counterfactual Explanations Can Be Manipulated*. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2106.02666>. Citado 3 vezes nas páginas 72, 93 e 103.

SMUHA, N. Ethics Guidelines for Trustworthy AI. *European Commission*, p. 1–39, 2019. Disponível em: <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf>. Citado na página 50.

STEPHENSON, N. et al. Survey of Machine Learning Techniques in Drug Discovery. *Current Drug Metabolism*, v. 20, n. 3, p. 185–193, 2019. ISSN 13892002. Disponível em: <http://www.eurekaselect.com/article/92486>. Citado na página 35.

STERNIK, I. *La inteligencia que no piensa*. 2018. Disponível em: <https://www.pagina12.com.ar/109080-la-inteligencia-que-no-piensa>. Citado na página 27.

SWEENEY, L. Discrimination in online Ad delivery. *Communications of the ACM*, v. 56, n. 5, p. 44–54, 2013. ISSN 00010782. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/2460276.2460278>. Citado na página 64.

TECNOPOLÍTICA, S. A. *Inteligencia artificial, embarazo, Salta, Urtubey y la Fundación Conin - YouTube*. Salto Agencia Tecnopolítica, 2018. Disponível em: <https://www.youtube.com/watch?v=xyLuxPhCgBw>. Citado na página 21.

TOBORE, I. et al. Deep Learning Intervention for Health Care Challenges: Some Biomedical Domain Considerations. *JMIR mHealth and uHealth*, JMIR Publications Inc., v. 7, n. 8, aug 2019. ISSN 22915222. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6696854/>. Citado na página 15.

URTUBEY Y. *Urtubey y una insólita propuesta de "prever"embarazos adolescentes*. 2018. Disponível em: <https://www.diariodecuyo.com.ar/argentina/Urtubey-y-una-insolita-propuesta-de-prever-embarazos-adolescentes-20180411-0081.html>. Citado na página 23.

US FDA. *Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) - Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback*. [S.l.], 2019. 20 p. Disponível em: <https://www.fda.gov/media/122535/download>. Citado 2 vezes nas páginas 51 e 52.

VALENTE, J. *Riscos da inteligência artificial levantam alerta e suscitam respostas*. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2020-08/riscos-da-inteligencia-artificial-levantam-alerta-e-suscitam-respostas#>. Citado na página 27.

VARON, J.; PEÑA, P. Artificial intelligence and consent: a feminist anti-colonial critique. *Internet Policy Review*, Alexander von Humboldt Institute for Internet and Society, v. 10, n. 4, 2021. ISSN 21976775. Citado 2 vezes nas páginas 27 e 28.

VILONE, G.; LONGO, L. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, v. 3, n. 3, p. 615–661, 2021. ISSN 2504-4990. Disponível em: <https://www.mdpi.com/2504-4990/3/3/32>. Citado na página 33.

VILONE, G.; LONGO, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, Elsevier B.V., v. 76, n. June 2020, p. 89–106, 2021. ISSN 15662535. Disponível em: <https://doi.org/10.1016/j.inffus.2021.05.009>. Citado 2 vezes nas páginas 61 e 70.

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, p. 1–52, 2017. ISSN 1556-5068. Citado 2 vezes nas páginas 68 e 99.

WHO. *Adolescent pregnancy*. 2020. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/adolescent-pregnancy>. Citado 2 vezes nas páginas 20 e 24.

WHO. *Ethics and governance of artificial intelligence for health: WHO guidance*. [S.l.: s.n.], 2021. 165 p. ISBN 9789240012752. Citado 4 vezes nas páginas 20, 27, 28 e 35.

XAVIER, F. C. *A Estratégia Brasileira de Inteligência Artificial*. 2021. Disponível em: <https://mittechreview.com.br/a-estrategia-brasileira-de-inteligencia-artificial/>. Citado na página 50.

YAZLLE, M. E. H. D. Gravidez na adolescência. *Revista Brasileira de Ginecologia e Obstetrícia*, Federação Brasileira das Sociedades de Ginecologia e Obstetrícia, v. 28, n. 8, p. 443–445, aug 2006. ISSN 0100-7203. Disponível em: <http://www.scielo.br/j/rbgo/a/Y4NtJBwZGYcvCngcWzsgnXj/?lang=pt>. Citado na página 20.

ZHOU, Y.; RIBEIRO, M. T.; SHAH, J. ExSum: From Local Explanations to Model Understanding. 2022. Disponível em: <http://arxiv.org/abs/2205.00130>. Citado na página 75.

ZIEN, A.; KR, N. The Feature Importance Ranking Measure. p. 694–709, 2009. Disponível em: <https://link.springer.com/content/pdf/10.1007/978-3-642-04174-7_45.pdf>. Citado na página 88.

Anexos

ANEXO A – Variáveis utilizadas no modelo de prevenção de gravidez precoce

Variáveis utilizadas pelo Ministério da Primeira Infância de Salta (Argentina)

A seguir são listadas as variáveis que foram coletadas junto à população de Salta na Argentina para treinamento do modelo de predição de gravidez na adolescência. Segundo Ortiz Freuler e Iglesias (2018, p. 35), os dados foram fornecidos pelo Ministério da Primeira Infância via e-mail e podem ser acessados no arquivo (planilha *3. Full List of Variables, Tran*) disponível em http://bit.ly/ai-annex (acessado em 07/03/2022).

São ao todo 78 atributos, que podem ser agregados seguintes grupos: pessoais, educação, saúde, trabalho, moradia e família. As variáveis reunidas pelos autores estão disponíveis nos idiomas espanhol e inglês e são listadas nos quadros 5, 6, 7, 8, 9 e 10.

Quadro 5 – Gravidez precoce: dados pessoais

Dados pessoais	
Espanhol	**Inglês**
Código de Persona	*Person Code*
Código de Visita	*Visit Code*
Fecha de nacimiento	*Birthdate*
Etnia	*Ethnicity*
Estado civil	*Civil status*
Nacionalidad	*nationality*
País origen	*Country of origin*
Relación con el jefe de hogar	*Relationship with the head of household*

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

Registros de registro na escola Nível educacional máximo alcançado É analfabeto Participe do estabelecimento educacional Quanto tempo não participa do estabelecimento educacional

Quadro 6 – Gravidez precoce: dados sobre a educação

Dados sobre a educação		
Espanhol	**Inglês**	**Português**
Registra inscripción en escuela	*Register enrollment in school*	Possui inscrição escolar
Máximo nivel educativo alcanzado	*Maximum educational level achieved*	Nível educacional máximo alcançado
Es analfabeto	*He is illiterate*	É analfabeto
Concurre a establecimiento educativo	*Concurre to educational establishment*	Frequenta um estabelecimento de ensino
Cuánto tiempo hace que no concurre a establecimiento educativo	*How long have you been absent from an educational establishment?*	Há quanto tempo você está ausente de um estabelecimento de ensino?

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

Quadro 7 – Gravidez precoce: dados sobre a saúde

Dados sobre a saúde	
Espanhol	**Inglês**
Discapacitado (Si/No)	*Disabled (Yes / No)*
Cantidad de discapacidades	*Number of disabilities*
Tipo de discapacidad	*Type of disability*
Posee certificado de discapacidad	*Has a disability certificate*
Lo asiste acompañante terapéutico	*He is accompanied by a therapeutic*
Cobertura médica	*Medical coverage*
Posee cobertura médica	*It has medical coverage*
Tiene Enfermedades Crónicas	*Have Chronic Diseases*
Cantidad de Enfermedades Crónicas	*Number of Chronic Diseases*
Tiene Enfermedades Agudas	*Have Acute Diseases*
Cantidad de Enfermedades Agudas	*Number of Acute Diseases*
Tuvo embarazos anteriores	*Had previous pregnancies*

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

Quadro 8 – Gravidez precoce: dados sobre trabalho

Dados sobre trabalho	
Espanhol	**Inglês**
Cantidad de Planes Sociales	*Number of Social Plans*
Plan Social	*Social Plan*
Tiene Oficios	*It has Trades*
Trabaja algún oficio	*Work some trade*
Trabaja	*Work*
Situación Laboral	*Employment situation*
Ámbito de Ocupación	*Scope of Employment*
Trabajo no formal	*Non-formal work*
Motivo por el que No Trabaja	*Reason for not working*

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

Quadro 9 – Gravidez precoce: dados sobre moradia

Dados sobre a moradia	
Espanhol	**Inglês**
Cantidad de personas que habitan la vivienda	*Number of people living in the house*
Cantidad de Varones que habitan la Vivienda	*Number of Males living in Housing*
Cantidad de Mujeres que habitan la Vivienda	*Number of Women living in Housing*
Cantidad de Analfabetos que habitan la Vivienda	*Number of Illiterates that live in Housing*
Cantidad de Discapacitados que habitan la Vivienda	*Number of Disabled people living in Housing*
Tipo de Vivienda	*Type of Housing*
Hacinamiento	*Overcrowding*
Material de piso	*Floor material*
Material de pared	*Wall material*
Material de Techo	*Roof Material*
Tiene Baño	*Has a bathrom*
Cantidad de Baños	*Amount of Bathrooms*
Ubicación del Baño	*Bathroom Location*
Tiene Agua Caliente en el baño	*It has hot water in the bathroom*
Tipo de Desagüe	*Type of Drain*
Tiene baño con botón o cadena	*It has a bath with button or chain*
Tiene inodoro con mochila	*It has a toilet with a backpack*
Fuente de Electricidad	*Source of Electricity*
Fuente de Gas	*Gas Source*
Fuente de Residuos	*Waste Source*

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

Quadro 10 – Gravidez precoce: dados sobre família

Dados sobre a família	
Espanhol	**Inglês**
Edad de padre	*Father's age*
Padre trabaja	*Father works*
Tipo de trabajo de padre	*Type of father work*
Estado civil de padre	*Marital status of father*
País de origen de padre	*Country of origin of father*
Etnia de padre	*Father's ethnic group*
Máximo nivel educativo alcanzado por el padre	*Maximum educational level reached by the father*
Edad de jefe de hogar	*Age of head of household*
Jefe de hogar trabaja	*Head of household works*
Tipo de trabajo de jefe de hogar	*Type of work of head of household*
Estado civil de jefe de hogar	*Marital status of head of household*
País de origen de jefe de hogar	*Country of origin of head of household*
Etnia de jefe de hogar	*Ethnicity of head of household*
Máximo nivel educativo alcanzado por el jefe de hogar	*Maximum educational level achieved by the head of household*
Edad de madre	*Mother's age*
Madre trabaja	*Mother works*
Tipo de trabajo de madre	*Type of mother's work*
Estado civil de madre	*Marital status of mother*
País de origen de madre	*Country of origin of mother*
Etnia de madre	*Ethnicity of mother*
Máximo nivel educativo alcanzado por el madre	*Maximum educational level reached by the mother*
Madre tuvo embarazo siendo adolescente	*Mother had pregnancy as a teenager*
Alguna hermana tuvo embarazo siendo adolescente	*Some sister had pregnancy as a teenager*

Fonte: Ortiz Freuler e Iglesias (2018, p. 35).

ANEXO B – Cópia do repositório GitHub contendo alguns passos metodológicos da criação do modelo de predição de gravidez precoce

Este anexo é uma cópia capturada em 13/09/2020 do repositório GitHub de Facundo Davancens, funcionário da Microsoft na Argentina, sobre alguns passos metodológicos para a construção do modelo de predição de gravidez precoce. As imagens que integravam este documento não estavam disponíveis no momento de criação

Predicción de Embarazo Adolescente con Machine Learning

En este caso de estudio, me gustaría contarte cómo hicimos para detectar jóvenes adolescentes en riesgo de quedar embarazada utilizando técnicas de Machine Learning.

Puedes usar este caso como una guía para crear tu propio modelo, o simplemente leerlo para conocer la lógica y secuencia de pasos que seguimos para llegar a nuestro resultado. En la misma línea y si es la primera vez trabajas con Machine Learning, te recomiendo que leas el caso de predicción de deserción escolar documentado por mi colega Marcelo Felman con quien tuvimos el agrado de trabajar junto al Ministerio de Primera Infancia en Salta.

Resumen

En colaboración con el Ministerio de Primera Infancia del Gobierno Provincial de Salta, definimos como objetivo utilizar inteligencia artificial para identificar aquellas adolescentes con mayor riesgo de quedar embarazada.

Afortunadamente, tuvimos acceso a un amplio espectro de datos. Utilizamos un *dataset* de más de 200.000 residentes de la ciudad de Salta con más de 12.000 mujeres de entre 10 y 19 años , Argentina. **Estos datos no contienen información personal identificable sobre las personas, tal como reconoce habeas data**.

A través de las herramientas Azure Machine Learning y SQL Server 2016, logramos crear diferentes modelos predictivos que permiten detectar hasta el 90% de los casos en riesgo de embarazo adolescente.

Contexto

Salta es una de las ciudades más pobladas de Argentina y capital de la provincia con igual nombre.

Con una población que supera los 500.000 habitantes, su Ministerio de Primera Infancia tiene por misión erradicar la pobreza en la provincia. Con este objetivo en mente, el desarrollo de las capacidades de los individuos es algo fundamental. El Ministerio de Primera Infancia propone actuar de manera proactiva identificando con Inteligencia Artificial aquellas jóvenes en riesgo de quedar embarazada durante su adolescencia.

La duración del proyecto fueron dos semanas, en las cuales iniciamos desde la exploración del dominio hasta la publicación de un servicio web predictivo.

Software y herramientas

- Azure Machine Learning
- SQL Server 2016

Fases del proyecto

- Exploración del dominio
- Preparación de los datos
- Creación de modelos
- Integración

Exploración del dominio

Antes de empezar, debemos comprender el dominio en el cual estamos trabajando. Si bien lo ideal es contar con un experto en el dominio, no siempre lo tendremos a disposición. Para esto, es importante tener en cuenta diferentes técnicas así como también ejercitar nuestro sentido común.

Para comenzar a visualizar los datos, hemos utilizado un motor de bases de datos: SQL Server. De esta forma puedo escribir consultas SQL fácilmente y además exportar como CSV o inclusive conectar con Azure Machine Learning.

Para este proyecto, contamos con una tabla principal la cual contiene información de las personas: edad, etnia, país de origen, discapacidad, cantidad de personas con las que vive, si tiene agua caliente en el baño, barrio/zona en donde residen, si tuvo embarazo adolescente y si el jefe de hogar abandonó los estudios. A su vez, contiene información referencial (*foreign keys*) hacia sus padres o jefes de hogar.

Una práctica que encuentramos conveniente es unir todos los datos en una única tabla o proyección a través de la cláusula *JOIN*. De esta forma, podrás portarlo de manera más simple a Azure Machine Learning.

> **Tip:** Unir todos los datos en una única tabla o proyección en lugar de lidiar con muchos conjuntos de datos.

Ejemplo #1

Supongamos que sólo quiero conocer la relación entre las adolescentes embarazadas y saber si sus madres abandonaron los estudios, haríamos algo así:

```
SELECT
        chica.Embarazo AS EmbarazadaAdolescente,
        madre.AbandonoEstudios AS MadreAbandono,
        COUNT(*) AS Interseccion
FROM
        PersonasSaltaCapital chica
INNER JOIN
        PersonasSaltaCapital madre
ON
        chica.CodMama = madre.CodPersona
WHERE
                chica.Embarazo = 'SI'
        AND chica.Edad <= 19
        AND chica.Sexo='Femenino'
GROUP BY
        chica.Embarazo,
        madre.AbandonoEstudios
ORDER BY
```

Esto es importante saberlo, ya que en la fase inicial lo haremos muchas veces.

Pruebas simples

Para ir familiarizándote con el conjunto de datos, es bueno que ejecutes algunas consultas. Estas corrí yo como ejemplo:

- Cantidad total de personas
- Cantidad de jóvenes embarazadas versus las que no (menores de 20 años)
- Cantidad de jóvenes embarazadas que abandonaron los estuidos versus las que no
- Tasa de embarazo por zona en donde viven
- Cantidad de jóvenes embarazadas agrupadas por país de origen
- Cantidad de jóvenes embarazadas por etnia

> **Tip:** Invierte todo el tiempo que creas necesario para entender las relaciones entre los datos. 2 o 3 días enfocado en esto puede ser un tiempo razonable (aunque parezca mucho), dependiendo del dominio.

Ejemplo #2

Este es un ejemplo para calcular la cantidad total de jóvenes embarazadas:

```
SELECT COUNT(*)
FROM Personas
WHERE Embarazo = 'SI'
        AND Edad <= 19
        AND Sexo='Femenino'
```

Este es un ejemplo para calcular los embarazos por zona:

```
SELECT TOP 10
        Barrio,
        COUNT(*) AS EmbarazadasAdolescentes,
        (SELECT COUNT(*) FROM PersonasSaltaCapital psc2 WHERE psc2.Barrio
        CAST(COUNT(*) AS FLOAT) / CAST((SELECT COUNT(*) FROM PersonasSalta
FROM PersonasSaltaCapital psc
WHERE
        psc.Embarazo = 'SI'
        AND psc.Edad <= 19
        AND psc.Sexo='Femenino'
GROUP BY
        Barrio
ORDER BY
        Porcentaje DESC
```

Aquí, obtuvimos resultados similares a los siguientes (muestro el TOP 10 para simplificar):

Barrio	Embarazadas Adolescentes	TotalBarrio	Porcentaje
Virgen de Urkupina	2	28	7,14285714285714
Finca La Paz	1	17	5,88235294117647
16 de Septiembre	1	21	4,76190476190476
Isla Soledad	2	43	4,65116279069767
El Circulo VII	1	23	4,34782608695652
Las Costas	3	88	3,40909090909091
Martin M. de Güemes	2	67	2,98507462686567
Manantial del Sur	3	118	2,54237288135593
Santa Rita S	1	40	2,5
San Isidro	2	86	2,32558139534884

Como puedes ver, distintas zonas tienen distintas tasas de embarazo. Esto pueda llevarnos a pensar que la zona donde una adolescente resida, tenga que ver con su probabilidad de quedar embarazada. Por ahora, es sólo una hipótesis que luego probaremos.

> **Tip:** Seguro encuentres mejores maneras de escribir estas consultas. No te preocupes, la performance aquí no es importante.

Iteración

A estas alturas, empezarás a darte cuenta de ciertas variables que pueden llegar a ser o no relevantes en nuestro modelo. Proyécatalas todas y luego en Azure Machine Learning validaremos cuáles son las que sirven.

Para nuestro modelo, utilizaremos inicialmente las siguientes variables:

- Edad
- Barrio donde reside
- Etnia

- Pais de Origen
- Tipo de discapacidad, si es que tiene
- Tiene Agua Caliente en el Baño
- Cantidad de personas con quien vive
- Jefe de hogar abandonó estudios

Luego iteraremos para entender si estas variables nos sirven o no.

Preparación de los datos

Ahora que sabemos qué datos queremos utilizar, debemos prepararlos para ir hacia Azure Machine Learning. Estando en SQL Server, una manera simple de hacerlo es guardando los resultados de nuestra consulta como CSV.

Esto nos generará un archivo de tipo .rpt, al cual simplemente cambiar su extensión a .csv y abrirlo como tal.

> **Tip:** Los archivos CSV pueden abrirse con Excel. Si te hace sentir más cómodo, chequealo antes de seguir con el próximo paso.

Ahora, procedemos a ingresar al Azure Machine Learning Studio. Si no tienes una cuenta, puedes crearla de forma gratuita allí mismo.

Al ingresar, hacemos click en *New* y luego en *Dataset*


Subir el conjunto de datos

Subimos nuestro conjunto de datos, y al cabo de unos segundos estará en Azure.

Ahora puedes crear un nuevo experimento haciendo click en *+NEW* > *Blank Experiment*

> **Tip:** Puedes ponerle el nombre que quieras haciendo click en el nombre por defecto.

Ya subido el dataset, el mismo aparecerá en la solapa *My datasets* y lo podrás utilizar inmediatamente. Para esto, simplemente lo arrastras a la ventana principal.


Nuevo experimento

Si quieres visualizarlo, puedes hacer click en la salida del módulo y otro click en *visualize*.


Visualizar conjunto de datos

Deberías ver una pantalla similar a esta, pero con tus propios datos. Si haces click en cada una de las columnas, podrás ver la distribución junto con un histograma a la derecha.

Creación de modelos

Ahora empezaremos a crear nuestros modelos predictivos. Como mencionamos anteriormente, lo que estamos buscando es predecir *si una joven tiene o no un embarazo durante la adolescencia*. En otras palabras, es una clasificación binaria (de dos clases). Para más información sobre clasificación binaria y otros tipos de problemas, puedes ver aquí.

Para resolver este problema, existen diferentes algoritmos. Recuerda que los diferentes algoritmos son simplemente distintas formas de abordar a un resultado. Algunos llegan a mejores resultados, pero esto no está garantizado. Si quieres ver el listado y ventajas de cada uno de los algoritmos, lo puedes ver aquí. En este caso comenzamos utilizando un *Two-class Boosted Decision Tree*.

Para identificar la columna a predecir utilizaremos *Edit Metadata*. Debemos conectar la salida del dataset con la entrada de *Edit Metadata*. Usa el selector de columna a la derecha para elegir el campo.

El dataset con el cual estamos trabajando tiene un 7% de jóvenes mujeres que tienen o tuvieron un embarazo adolescente. Si lo pensamos en frío, con tener un modelo que diga no ya estariamos cubriendo el 93% de los casos, pero lo que corresponde es balancear el conjunto de datos como se explica a continuación.

Balancear el conjunto de datos

Los conjuntos de datos desbalanceados son un problema típico. Esta situación puede generar un balanceo que favorezca el escenario mayoritario, es decir el 'NO curso un embarazo adolescente'. Para ello, podemos aplicar dos técnicas distintas:

- *Undersampling*: tomar menos casos del escenario mayoritario a fin de reducir las ocurrencias.
- *Oversampling*: aumentar o simular más ocurrencias del caso minoritario.

En este caso, optamos por usar *Oversampling*. De esta manera, logramos un conjunto de datos más balanceado. Una manera simple y casi automática de lograr esto, es utilizando el módulo *SMOTE*, que significa *Synthetic Minority Oversampling* o bien sobre-muestreo sintético de minorías. No olvides aquí también editar metadata. Más información sobre SMOTE, aquí.

> **Tip:** En la solapa Propiedades de *SMOTE*, puedes ajustar el porcentaje de aumento de las ocurrencias minoritarias. Puedes jugar con este número hasta alcanzar un resultado que te sirva.

Realizar las predicciones y Evaluar el comportamiento del modelo

Utilizaremos el módulo *Cross-Validate Model* el cual toma como entrada la salida de *Two-Class Boosted Decision* y de *SMOTE*. Acto seguido, no olvidar seleccionar la columna que corresponda, en nuestro caso, la que indica si cursa o cursó un embarazo adolescente.

Por último evaluaremos el modelo utilizando el módulo *Evaluate model*, el cual tendrá conectada en su entrada la salida proveniente de *Cross-Validate Model*

Si damos *Run* visualizamos la salida de *Evaluate model*, podremos ver la siguiente gráfica:

Esta gráfica demuestra los casos que fueron correctamente identificados, respecto los que no. Esencialmente, el área debajo de la curva (*Area under the curve* o también *AUC*) debe ser lo mayor posible: no queremos dejar casos afuera.

A simple vista, nuestro modelo parece comportarse de una manera muy acertada: el **98,2%** de las veces está realizando una predicción correcta.

No obstante, si nos movemos hacia abajo veremos más métricas que definen el comportamiento de nuestro modelo predictivo.

Como podemos apreciar, lo que estamos haciendo es identificar cuatro casos distintos:

- Jóvenes mujeres que dijimos que tienen o tuvieron embarazo adolescente y efectivamente tuvieron un embarazo adolescente (verdadero positivo): 1516.

- Jóvenes mujeres que dijimos que NO tienen o tuvieron embarazo adolescente pero tuvieron un embarazo adolescente (falso negativo): 169.
- Jóvenes mujeres que dijimos que tienen o tuvieron embarazo adolescente pero NO tuvieron un embarazo adolescente (falso positivo): 72.
- Jóvenes mujeres que dijimos que NO tienen o tuvieron embarazo adolescente y efectivamente fue así (verdadero negativo): 11702.

Debemos ser cautelosos y evaluar estos puntos. Una buena pregunta para hacernos es cuál es el costo de cada escenario. En este caso, es mucho peor NO ayudar a una joven mujer en riesgo de quedar embarazada durante su adolescencia que ayudar por demás a una adolescente que NO quedará embarazada durante su adolescencia. En otras palabras, el costo de los falsos negativos supera el de los falsos positivos.

Como parte del proceso hemos experimentado modificando los valores de *SMOTE*, pero la mejor *sencibilidad* (o recall) que obtuvimos es de 0,9 que nos indica la proporción de eventos positivos identificados correctamente.

Integración

Creación del Web Service

Hemos llegado a la instancia donde estamos conformes sobre nuestro modelo, y queremos que sea consumido. La forma más simple será a través de un *Web Service REST*.

Para crear el servicio, debemos correr nuestro experimento (si es que no lo hicimos) y hacer click en el botón de *SET UP WEB SERVICE*.


Preparar el Web Service

Esto nos generará una animación y creará una pestaña con el servicio web.

> **Tip:** No te preocupes por hacer modificaciones en esta instancia.

Deberás darle *Run* al servicio nuevamente, que es como si fuera "compilar" nuestra nueva API.

Finalmente, aparecerá el botón *DEPLOY WEB SERVICE* sobre el cual haremos click. Al cabo de unos segundos, nuestro servicio web estará listo para consumir.


Desplegar el Web Service

Probar el Web Service

Una vez desplegado, verás la siguiente pantalla.

Notarás que tienes dos formas de utilizarlo:

- *Request/Response*: puedes generar la predicción para un único caso.
- *Batch execution*: puedes realizar muchas predicciones con un mismo pedido a la API.

> **Tip:** Te recomiendo arrancar por *request/response* si es tu primera vez.

Estos servicios pueden probarse directamente desde el portal. Para ello, puedes hacer click en el botón *TEST*. Te permitirá ingresar algunos campos, para finalmente darte una respuesta en formato *JSON*.

> **Tip:** También puedes probarlo integrándote con Excel. Es bastante simple y puede ahorrarte tiempo para ingresar atos.

De esta manera, tu servicio web predictivo ya es visible para el mundo exterior. ¡Felicitaciones!

Conclusiones

Machine Learning es todo un mundo diferente, y puede resultar complejo para quienes venimos del ámbito de desarrollo de software.

A través de un proceso iterativo de prueba y error, podemos ir acercándonos a una respuesta correcta y finalmente determinar si nuestro modelo es bueno o no.

En nuestro caso, logramos un modelo predictivo que identifica correctamente a aproximadamente el 90% de las jóvenes mujeres están en riesgo quedar embarazadas durante su adolescencia. Esta herramienta permite al Gobierno tomar decisiones en tiempo real, y ayudar a aquellos que más lo necesiten.

Agradecimientos

El mayor agradecimiento al Ministerio de Primera Infancia del Gobierno Provincial de Salta, quienes sin lugar a dudas quieren cambiar este mundo para el bien de todos y a Microsoft por posibilitarnos el hecho de ser parte de dicho cambio.

ANEXO C – Cooperação técnica entre o Ministério da Cidadania e a Microsoft do Brasil

Íntegra do acordo de cooperação técnica entre o Ministério da Cidadania e a Microsoft do Brasil LTDA.

ACORDO DE COOPERAÇÃO TÉCNICA Nº 47 /2019

PROCESSO Nº 71000.036620/2019-43

ACORDO DE COOPERAÇÃO TÉCNICA QUE ENTRE SI CELEBRAM A MINISTÉRIO DA CIDADANIA E A **E A MICROSOFT DO BRASIL IMPORTAÇÃO E COMÉRCIO DE SOFTWARE E VÍDEO GAMES LTDA**, VISANDO A CONSECUÇÃO DE PROVA DE CONCEITO PARA IMPLEMENTAR FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL QUE SUBSIDIEM MELHORIA DAS AÇÕES DO PROGRAMA CRIANÇA FELIZ.

O MINISTÉRIO DA CIDADANIA, com sede na Esplanada dos Ministérios, Bloco A, 7º andar, CEP 70054-906, Brasília/DF, inscrito no CNPJ/MF sob nº 05.526.783/0001-65 , doravante denominado MINISTÉRIO, representado neste ato por **OSMAR GASPARINI TERRA**, Ministro de Estado da Cidadania, , nomeado por Decreto Presidencial de 1º de janeiro de 2019, publicado no Diário Oficial da União- Edição Especial, de 1º de janeiro de 2019, na Seção 2, página 1, no exercício da atribuição que lhe confere a Lei nº 10.683, de 28 de maio de 2003, alterada pela Lei nº 10.869, de 13 de maio de 2004, e posteriormente pela Lei nº 13.341, de 29 de setembro de 2016; e a **MICROSOFT DO BRASIL IMPORTAÇÃO E COMÉRCIO DE SOFTWARE E VIDEO GAMES LTDA.**, sociedade limitada com sede na cidade de São Paulo, estado de São Paulo, na Avenida Presidente Juscelino Kubitschek, nº 1.909, conjunto 171, 17º andar da Torre Sul SP Corporate Towers, inscrita no CNPJ/MF sob nº 04.712.500/0001-07, doravante denominada MICROSOFT; representada neste ato por **RONAN TEIXEIRA DAMASCO**, residente e domiciliado em Brasília/DF, inscrito no CPF sob o número 287.351.451-53;

CONSIDERANDO que o **MINISTÉRIO** deseja desenvolver um trabalho de análise para o programa Criança Feliz, utilizando ferramentas tecnológicas de processamento de dados baseado em inteligência artificial como mecanismo de diagnóstico orientado a detectar situações de vulnerabilidades sociais como guia para formulação de políticas públicas preventivas e transformadoras. O MINISTÉRIO, como linha de ação permanente, destaca atenção primordial à primeira infância, uma vez que o investimento nesta camada da população aumenta as possibilidades de uma sociedade mais próspera.

CONSIDERANDO que a **MICROSOFT** é uma empresa de tecnologia com vasto *expertise* em serviços de Computação em Nuvem e Inteligência Artificial cuja missão é empoderar cada pessoa e organização a conquistar mais, promovendo a inclusão social e gerando progresso, em alinhamento com os objetivos de desenvolvimento sustentável da ONU, a fim de diminuir a desigualdade no acesso às habilidades digitais, em especial para jovens em situação de

risco, apoiando ações humanitárias para construir comunidades fortes com uso de tecnologia para fomentar o desenvolvimento das futuras gerações.

CONSIDERANDO que a MICROSOFT já desenvolveu projeto semelhante, tendo por referência os "considerandos" acima, com a PROVÍNCIA DE SALTA, na República Argentina, e pode se valer toda experiência e inteligência adquirida com o mesmo, pelo presente, é estabelecida uma cooperação para eventual desenvolvimento, adequação e uso de uma plataforma no Brasil.

CONSIDERANDO que o **MINISTÉRIO** e a **MICROSOFT** possuem uma visão comum que inclui ajudar a todas as pessoas, em especial crianças e adolescentes, ao redor do território brasileiro, para que possam atingir seu potencial pleno; e que a **MICROSOFT** pode cooperar com o compartilhamento de experiências e conhecimentos sobre o uso de tecnologias para tal processo;

RESOLVEM, celebrar o presente Acordo de Cooperação ("Acordo"), aplicando-se, no que couber a Lei nº 8.666, de 21 de junho de 1993, mediante as seguintes cláusulas e condições:

CLÁUSULA PRIMEIRA – DO OBJETO

1.1 O presente Acordo tem por objetivo estabelecer a cooperação técnica das partes para fim de consecução de prova de conceito para implementar ferramentas de inteligência artificial que subsidiem a melhoria do desenvolvimento dos direitos das crianças e adolescentes, bem como o fortalecimento de suas famílias e comunidades, através de ações do programa Criança Feliz.

1.2 A cooperação objetiva construir, em conjunto, uma solução que coleta de dados através de formulários eletrônicos e uso de ferramentas analíticas e de inteligência artificial sobre esses dados para subsidiar ações do programa Criança Feliz.

1.3 O presente Acordo se restringe à cooperação descrita acima e execução das respectivas atividades relacionadas, devendo as partes detalhar ao menos um Plano de Trabalho para definição de atividades relacionadas ao escopo ora definido, antes do início das atividades cooperadas.

1.4 Este Acordo não cria obrigação de entrega ou desenvolvimento de qualquer solução específica e tampouco estabelece qualquer obrigação futura de aquisição de produtos, serviços ou licenças por parte do **MINISTÉRIO**.

CLÁUSULA SEGUNDA – DA EXECUÇÃO/ ATRIBUIÇÕES

2.1 Cada parte designará responsável pelo acompanhamento e monitoramento da execução do pactuado no presente Acordo.

2.2 Constituem atribuições das partes:

a) receber em suas dependências o(s) responsável(is) indicado(s) pela outra parte para participar do desenvolvimento de atividades atinentes ao objeto do presente Acordo;
b) cumprir os encargos e obrigações estabelecidos neste acordo e eventuais Planos de Trabalho, assumindo cada parte os seus respectivos custos na execução e entrega da cooperação ora estabelecida;
c) cumprir com todas as leis anticorrupção aplicáveis, coibindo a prática de atos fraudulentos ou de corrupção, incluindo-se condutas que visem obter vantagem ilícita, em prejuízo alheio; e

d) levar ao conhecimento da outra parte, fato ou ocorrência que interfiram no andamento das atividades decorrentes deste Acordo.

2.3 O **MINISTÉRIO** deverá: (a) acompanhar e auxiliar o pessoal da **MICROSOFT** na execução das atividades; (b) fornecer acesso às suas instalações em que serão realizadas as atividades, conforme necessário; (c); disponibilizar a materiais necessários para a execução das atividades; (d) alocar os recursos humanos necessários para cumprir com as atividades relacionadas ao presente Acordo, e (e) compartilhar as informações relacionadas ao objeto da presente cooperação, de modo a viabilizar o desenvolvimento da solução referida no item 1.2 acima.

2.4 A **MICROSOFT**, no âmbito deste Acordo, deverá: (a) aportar recursos detalhados em Plano de Trabalho para compartilhar conhecimento técnico e atividades com o propósito do escopo acima estabelecido; (b) cooperar para o objeto a título de prova de conceito; (c) colocar à disposição do **MINISTÉRIO**, e a seus parceiros estratégicos, seus programas de responsabilidade social conforme as condições de elegibilidade dos mesmos; (d) prover as informações necessárias sobre as soluções e ferramentas tecnológicas da Microsoft com o propósito de alcançar os objetivos deste Acordo de Cooperação; (e) colaborar para a revisão de acordos específicos que sejam necessários para o lançamento dos projetos que sejam desenvolvidos no marco do presente Acordo de Cooperação.

CLÁUSULA TERCEIRA – DA PROPRIEDADE INTELECTUAL E DA CONFIDENCIALIDADE

3.1 As partes concordam que nenhuma das disposições do presente instrumento deverá ser interpretada como forma de licença ou cessão de direitos de propriedade intelectual por qualquer das partes. Com efeito, cada uma das Partes permanecerá a única e exclusiva titular de seus respectivos direitos de propriedade intelectual. Nesse sentido, as partes reconhecem que todo material, informação, conhecimento e item de propriedade intelectual da **MICROSOFT**, utilizados na execução deste Acordo, permanece como propriedade intelectual exclusiva da **MICROSOFT**, não sendo transferida por ocasião do presente Acordo, o que inclui, também, qualquer material desenvolvido parcialmente ou totalmente pela **MICROSOFT** em conexão com este Acordo.

3.2 Com o término deste Acordo, por qualquer motivo, cessa, de imediato, qualquer uso autorizado de bens de propriedade intelectual, realizado em razão e sob este Acordo, exceto se as partes estabelecerem especificamente em sentido contrário a possibilidade de seu uso, o que deverá constar de termo de encerramento deste Acordo.

3.3 As partes, neste ato, obrigam-se por si, seus representantes, prepostos, funcionários, colaboradores e/ou subcontratados a tratar com absoluto sigilo e confidencialidade toda e qualquer informação, dados, materiais, pormenores, documentos, especificações técnicas ou comerciais, inovações e aperfeiçoamentos, desenhos, projetos, procedimentos, manuais, nome, relação e/ou base de dados de clientes e/ou de fornecedores ("Informações Confidenciais") dos quais venham a ter conhecimento ou acesso, ou que lhes sejam confiados em razão deste Acordo, não podendo, em nenhuma hipótese, proceder à reprodução, demonstração, fornecimento, revelação e/ou divulgação, total ou parcial, de qualquer informação para terceiros sob qualquer forma e pretexto, tampouco utilizá-los em proveito próprio ou de terceiros para fins estranhos aos do presente Acordo. Rescindido ou findo o presente instrumento, as partes obrigam-se a restituir

todos os documentos a elas entregues e que contenham informações recebidas ou obtidas no período de vigência deste Acordo, salvo aquelas que por sua natureza devam ser, exclusiva e obrigatoriamente, mantidos pelas Partes como prova de suas obrigações, inclusive perante terceiros.

3.4 As obrigações de confidencialidade previstas nesta cláusula com relação às Informações Confidenciais não serão aplicáveis às seguintes hipóteses: (i) aquelas que a qualquer tempo se tornem de domínio público, sejam ou tenham sido levadas a público, sem que fique configurada infração contratual; (ii) as informações sejam conhecidas por uma das partes antes de sua divulgação pela outra Parte ou que tenha sido independentemente desenvolvida pelos representantes da respectiva parte, sem que estes tenham tido acesso às Informações Confidenciais; (iii) as informações sejam divulgadas, de boa-fé, por terceiro legalmente legitimado e/ou intitulado para tanto; e (iv) a revelação das informações seja requerida por lei, ordem judicial e/ou determinação de órgão/agência governamental devidamente amparado em dispositivo legal.

3.5 A classificação de qualquer informação como sigilosa obedecerá ao procedimento previsto na Lei nº 12.527/2011.

CLÁUSULA QUARTA – DOS RECURSOS FINANCEIROS

4.1 O presente Acordo não implicará repasse de recursos financeiros entre as partes. Cada uma das partes assumirá seus próprios custos em decorrência dos recursos alocados na execução do escopo e suas atribuições, inexistindo qualquer obrigatoriedade prévia de assunção de obrigações a partir dos seus resultados.

4.2 Eventuais repasses de recursos financeiros ou de bens que se fizerem necessários deverão ser estabelecidos em instrumentos próprios com obediência aos princípios da administração pública, à Lei de Licitação e às normas aplicáveis.

CLÁUSULA QUINTA – DO ADITAMENTO

5.1 Eventual alteração a este Acordo deverá ser feita por escrito, mediante termo aditivo a ser assinado pelas partes.

CLÁUSULA SEXTA – DA VIGÊNCIA E DO TÉRMINO

6.1 O presente Acordo vigorará pelo prazo de 6 (seis) meses, a contar da data de sua assinatura, somente sendo prorrogado mediante termo aditivo.

6.2 O presente instrumento poderá ser resolvido no caso de descumprimento de qualquer de suas obrigações, depois de previamente notificado a parte em situação de descumprimento, e caso o mesmo não tenha sanado a questão no prazo de 15 (quinze) dias a contar da data do recebimento da notificação. Implicação na resolução automática do presente Acordo qualquer descumprimento relacionado a questões de propriedade intelectual, confidencialidade ou na inobservância das leis anticorrupção.

6.3 O presente instrumento poderá ser resilido, por iniciativa de qualquer parte, a qualquer tempo, sem qualquer ônus, encargos ou penalidades, mediante denúncia a ser notificada por escrito.

CLÁUSULA SÉTIMA – DA RESPONSABILIDADE

7.1 As finalidades estabelecidas neste Acordo não geram responsabilidades de qualquer natureza para as partes no caso de falha em seu atingimento ou consecução, renunciando as partes, expressamente, a qualquer direito de reivindicar quaisquer danos nesse sentido. As finalidades previstas no presente Acordo não implicam, sob nenhuma circunstância, obrigações vinculantes e não geram qualquer tipo de indenização em juízo ou fora dele. A **MICROSOFT** não garante nem assume responsabilidade por perdas e danos de qualquer tipo que possam decorrer, de forma exemplificativa: (i) da adequação das atividades previstas neste Acordo aos propósitos do **MINISTÉRIO** ou pela entrega de qualquer solução efetiva; e (ii) pela qualidade, legalidade, confiabilidade e utilidade de serviços, informações, dados, arquivos, produtos e qualquer tipo de material utilizados pelas outras partes ou por terceiros.

7.2 O presente Acordo não substitui qualquer outro acordo ou contrato eventualmente existente entre as partes, tampouco rege a licença de uso de qualquer software ou produto da **MICROSOFT**, que são governados pelos contratos específicos e aplicáveis conforme o caso.

CLÁUSULA OITAVA– DAS DISPOSIÇÕES GERAIS

8.1 O presente instrumento constitui o acordo integral entre as partícipes, substituindo quaisquer entendimentos anteriores, verbais ou por escrito, somente podendo ser alterado mediante termo aditivo.

8.2 Os direitos e obrigações decorrentes do presente Acordo não poderão ser cedidos ou transferidos para terceiros sem a prévia e expressa autorização dos demais partícipes.

8.3 Se qualquer das partícipes deixar de exercer, à época, direito decorrente deste Acordo, tal ato não representará renúncia ou novação, devendo ser interpretado como mera liberalidade, podendo ser exercido a qualquer tempo, a não ser que as partícipes disponham expressamente contrário.

8.4 As partes conduzirão suas atividades em seus próprios nomes e serão separadamente responsáveis pelos atos e conduta de seus empregados e agentes.

8.5 Não será considerado descumprimento contratual eventual descumprimento ocasionado por motivo de força maior ou caso fortuito.

8.6 As partes reconhecem que o presente instrumento foi elaborado dentro dos mais rígidos princípios da boa-fé e da probidade.

8.7 Nenhuma das disposições do presente Acordo deve ser interpretada como impedimento para que a MICROSOFT coopere ou celebre contrato com qualquer outra pessoa ou entidade, bem como desenvolva, licencie, venda, distribua ou disponibilize a qualquer outra pessoa ou entidade, de outra forma, quaisquer informações, serviços, produtos ou materiais de sua propriedade, licenciados ou controlados pela MICROSOFT. Da mesma forma, o **MINISTÉRIO**, pelas disposições constantes deste Acordo, não estará impedido de celebrar contrato ou cooperar com qualquer pessoa ou entidade, assim como licenciar, contratar ou adquirir, de outra forma, quaisquer informações, serviços, produtos ou materiais de outra pessoa ou entidade. Em suma o presente Acordo não estabelece qualquer relação de exclusividade em relação ao seu objeto e não afeta a independência das partes no estabelecimento de cooperação com outras empresas, entidades e/ou organizações.

8.8 Aplicam-se ao presente Acordo as leis brasileiras. Fica eleito o foro da Capital Federal do Brasil, Brasília, para dirimir questões ou dúvidas oriundas do presente Acordo, renunciando as partes a qualquer outro, por mais privilegiado que seja.

Por estarem justas e contratadas, as partes firmam o presente instrumento em 2 (duas) vias de igual teor e forma, na presença das 2 (duas) testemunhas infra-assinadas.

Brasília, 23 de setembro de 2019.

OSMAR GASPARINI TERRA
MINISTÉRIO DA CIDADANIA

RONAN TEIXEIRA DAMASCO
MICROSOFT DO BRASIL IMPORTAÇÃO E COMÉRCIO
DE SOFTWARE E VÍDEO GAMES LTDA.

Testemunhas:

1.
Nome:
CPF/MF:

2.
Nome:
CPF/MF: 343181007-15

ANEXO AO ACORDO DE COOPERAÇÃO TÉCNICA

PLANO DE TRABALHO

COORDENAÇÃO DA EXECUÇÃO RECURSOS FINANCEIROS

MINISTÉRIO DA CIDADANIA- MC	(Sem repasses)
MICROSOFT DO BRASIL IMPORTAÇÃO E COMÉRCIO DE SOFTWARE E VÍDEO GAMES LTDA.	

DESCRIÇÃO DO OBJETO DO ACORDO DE COOPERAÇÃO:

Estabelecer a cooperação técnica das partes para fim de consecução de prova de conceito para implementar ferramentas de inteligência artificial que subsidiem a melhoria do desenvolvimento dos direitos das crianças e adolescentes, bem como o fortalecimento de suas famílias e comunidades, através de ações do Programa Criança Feliz.

DESCRIÇÃO DAS METAS E PARÂMETROS PARA SUA AFERIÇÃO:

Construir, em conjunto, uma solução que coleta de dados através de formulários eletrônicos e uso de ferramentas analíticas e de inteligência artificial sobre esses dados para subsidiar ações do programa Criança Feliz. As metas serão aferidas com análise e avaliação dos resultados sobre a utilização das ferramentas.

Previsão de início do objeto	Data da assinatura
Previsão de fim da execução	Fevereiro/2020

DESCRIÇÃO DAS ATIVIDADES E DA FORMA DE EXECUÇÃO:

O MINISTÉRIO deverá: (a) acompanhar e auxiliar o pessoal da MICROSOFT na execução das atividades; (b) fornecer acesso às suas instalações em que serão realizadas as atividades, conforme necessário; (c); disponibilizar a materiais necessários para a execução das atividades; (d) alocar os recursos humanos necessários para cumprir com as atividades relacionadas ao presente Acordo, e (e) compartilhar as informações relacionadas ao objeto da presente cooperação, de modo a viabilizar o desenvolvimento da solução referida no item 1.2 acima.

A MICROSOFT deverá: (a) aportar recursos detalhados em Plano de Trabalho para compartilhar conhecimento técnico e atividades com o propósito do escopo acima estabelecido; (b) cooperar para o objeto a título de prova de conceito; (c) colocar à disposição do MINISTÉRIO, e a seus parceiros estratégicos, seus programas de responsabilidade social conforme as condições de elegibilidade dos mesmos; (d) prover as informações necessárias sobre as soluções e ferramentas tecnológicas da Microsoft com o propósito de alcançar os objetivos deste Acordo de Cooperação; (e) colaborar para a revisão de acordos específicos que sejam necessários para o lançamento dos projetos que sejam desenvolvidos no marco do presente Acordo de Cooperação.

CRONOGRAMA DE EXECUÇÃO

Etapa	Responsável	Início	Término
Definição do município para realização do piloto	Ministério da Cidadania	Setembro/ 2019	Setembro/2019
Adequação da plataforma para utilização no Brasil	Microsoft	Setembro/ 2019	Setembro/ 2019
Capacitação da equipe nacional do Programa Criança Feliz para utilização das ferramentas e realização de ajustes necessários para o piloto – Brasília/DF	Microsoft	Setembro/ 2019	Setembro/ 2019
Capacitação da equipe municipal do Programa Criança Feliz para utilização das ferramentas – município piloto	Microsoft	Setembro/ 2019	Setembro/ 2019
Coleta de dados em campo pela equipe municipal do Programa Criança Feliz com a utilização das ferramentas e instrumentos - município piloto	Microsoft/Ministério da Cidadania	Setembro/ 2019	Outubro/2019
Análise dos resultados	Microsoft/Ministério da Cidadania	Outubro/2019	Novembro/2019
Avaliação dos Resultados	Microsoft/Ministério da Cidadania	Novembro/2019	Dezembro/2019
Finalização do Projeto	Microsoft/Ministério da Cidadania	Janeiro/2020	Fevereiro/2020

Este documento foi digitado e diagramado utilizando as tecnologias TEX, LATEX, TEXLive e TEXMaker; com estilos providos pela classe abnTEX2.

Gerado em 6 de setembro de 2022.