Universidade Estadual de Campinas
Instituto de Computação

Projeto de Mestrado: Proposta de arquitetura para soluções de IA envolvendo métricas de Fairness e AI Explainability

Aluno: Thales Eduardo Nazatto
Orientadora: Cecília Mary Fischer Rubira
Co-Orientador: Leonardo Montecchi

Resumo

Este projeto de Mestrado possui o objetivo de contribuir com o tema de Inteligência Artificial (IA) do ponto de vista da Engenharia de Software, visto que o uso de IA envolvendo grandes volumes de dados vem crescendo conforme nossa sociedade migra processos manuais de trabalho para soluções digitais e necessita de tomadas de decisão mais rápidas e assertivas. Entretanto, as métricas usadas inicialmente para definir a eficácia de um algoritmo se mostraram limitadas devido a barreiras éticas e legais, resultando em vieses que refletem a sociedade de maneira que não era esperada pelos desenvolvedores da solução. Desta forma, a elaboração de diretrizes arquiteturais para uma aplicação de IA incluir conceitos de Explainability AI e métricas de Fairness pode ajudar neste problema. Desenvolvendo componentes, documentando métodos e ferramentas, e realizando estudos de caso, é esperado que este problema seja controlado e facilite o desenvolvimento de softwares com o uso responsável de dados.

Abstract

This Master's Degree project aims to contribute to the theme of Artificial Intelligence (AI) from the Software Engineering point of view, since the use of AI involving large volumes of data has been growing as our society migrates manual work processes to digital solutions and needs faster and more assertive decision making. However, the metrics used initially to define the effectiveness of an algorithm proved to be limited due to ethical and legal barriers, resulting in biases that reflect society in a way that was not expected by the solution developers. Thus, the elaboration of architectural guidelines for an AI application including concepts of Explainability AI and Fairness metrics can help with this problem. By developing components, documenting methods and tools, and making case studies, it is expected that this problem will be controlled and facilitate software development with the responsible use of data.

1 Introdução

Técnicas de Inteligência Artificial e Aprendizado de Máquina já são utilizadas há bastante tempo no ramo da Computação. Ramos como robótica e jogos são grandes exemplos, dada a necessidade nos mesmos de automatizar comportamentos que seriam tidos como triviais para um ser humano. Entretanto, nos últimos anos ocorreu um crescimento no uso dessas tecnologias em aplicações tradicionais, com previsão de US$ 57 bilhões em investimentos em 2021, 480% maior em relação a 2017 [4]. No Brasil, o número de empresas de IA aumentou de 120 em 2018 para 206 em 2020 [5].

Isso se mostra possível devido a grande quantidade de dados processada diariamente pelas empresas, que coletam estatísticas toda vez que um usuário acessa suas aplicações. Com esses dados, podem traçar diferentes perfis e usar soluções de IA para ter tomadas de decisão mais assertivas com o objetivo de melhorar a experiência de usuário e corrigir problemas. Porém, muitas dessas soluções foram projetadas sem pensar em governança de dados como requisito de projeto, e se mostram ineficientes quando ela é tratada em consideração.

Governança de dados é um tema que entrou em evidência recentemente em países como o Brasil: Iniciativas como a LGPD - Lei Geral de Proteção de Dados [6], de 14 de agosto de 2018, mostram como as aplicações e seus dados possuem cada vez mais influência na sociedade moderna. E nesse ponto muitas aplicações de IA falham: muitas implementações foram implementadas como *black boxes*, onde o determinante para estabelecer a confiança no modelo implementado é sua entrada e sua saída.

Um outro efeito colateral dessa estratégia é a exposição de vieses que, embora sejam vistos como não-intencionais pelos desenvolvedores por ter a possibilidade de ser um *outlier* no modelo treinado, refletem preconceitos escancarados da sociedade atual. Uma entrada de dados enviesada resulta em um algoritmo que realiza discriminações em sua classificação [10], e uma vez que as métricas utilizadas para medir a qualidade de um modelo *black box* são geralmente baseadas em acurácia, precisão e recall, discriminações não são facilmente percebidas por tais métricas. Ao mesmo tempo, um modelo *black box* pode ter uma alta dependência de poucos dados, determinando problemas de acoplamento. Para resolver este problema, é possível que a criação de um novo modelo seja uma melhor opção que realizar a correção em apenas parte dele.

Devido a esses tipos de problemas, o termo *Explainable AI* (XAI) ganha força para envolver o desenvolvimento de uma IA que seja acurada e simultaneamente transparente. Como IA possui diversos tipos de métodos diferentes para enquadrar diversos tipos de dados, o mesmo acaba se aplicando em Explainable AI, podendo enquadrar em diversos tipos de dados [22], ou dados específicos como imagens [13] e tabelas [19]. No mesmo tema, é possível estabelecer métricas para determinar o quão o modelo está preparado para dados sensíveis [9], termo que é conhecido como *Fairness*.

Como o objetivo em XAI é fazer com que os resultados alcançados pela solução de IA sejam compreendidos por humanos, é possível considerar este fato como requisito no design de uma solução de IA, fazendo com que a mesma seja reusável e testável. O objetivo deste projeto de mestrado é estabelecer uma arquitetura para desenvolvimento de uma solução já baseada nestes conceitos, permitindo a criação de um algoritmo confiável e capaz de ser mantido por vários desenvolvedores.

O restante desse documento se organiza da seguinte forma: a Seção 2 descreve os conceitos que irão ser abordados neste projeto; a Seção 3 mostra trabalhos semelhantes; a Seção 4 enumera as resultados esperados desse projeto; a Seção 5 discute desafios a serem enfrentados e, finalmente, a Seção 6 estabelece o cronograma do projeto.

2 Conceitos Abordados

2.1 AI Explainability

AI Explainability é um conceito em IA que propõe a criação de um conjunto de técnicas de Aprendizado de Máquina (ou Machine Learning/ML) que produz modelos mais explicáveis, mantendo qualidade em suas métricas, e permite que os humanos entendam, confiem e gerenciem aplicações baseadas em IA. XAI também absorve conceitos das Ciências Sociais e considera a psicologia da explicação [7]. Um algoritmo de ML explicável precisa não apenas mostrar tomadas de decisão, mas também mostrar o processo que o levou ao tomar tal decisão, de modo que seja compreensível e transparente para humanos.

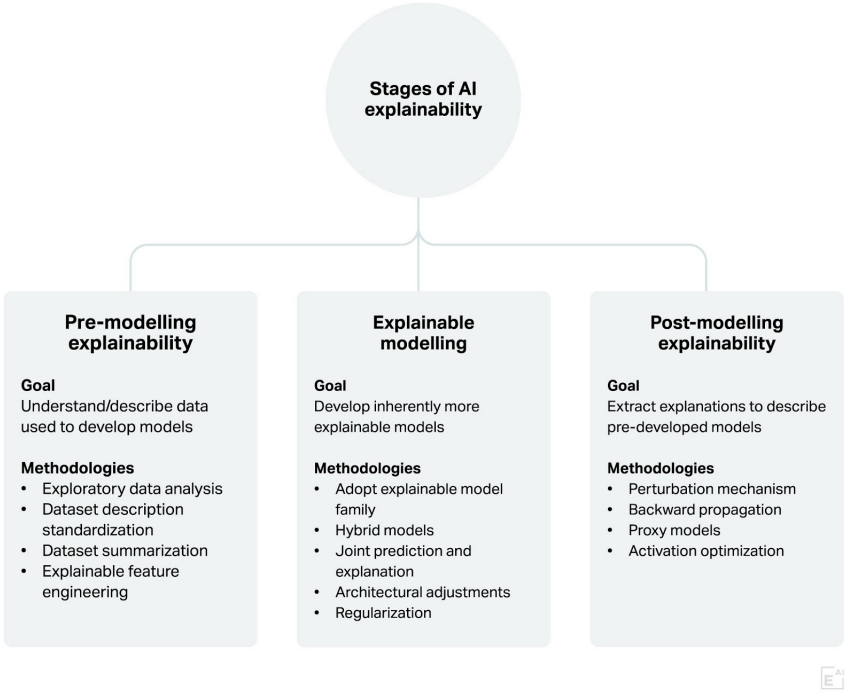

Figura 1: Fases de um processo baseado em IA explicável.

Conforme ilustrado na Figura 1, o um processo baseado em IA explicável pode ser classificado em 3 fases:

- ***Pre-Modelling Explainability*:** Esta fase se caracteriza por obter explicações no conjunto de dados usado para treino e validação, tendo como principal motivação o fato do comportamento de um modelo depender muito dos dados que o alimentam. Análises e visualizações dos dados, ou mesmo através de documentações do conjunto de dados se enquadram neste estágio. [18]

- ***Explainable Modelling*:** Esta fase se caracteriza por obter explicações por modelos considerados transparentes (ver adiante). Ao adotar modelos deste tipo, ele já pode ser considerado como explicável por um humano. Também é possível obter a explicação através de regularizadores, ou adotar abordagens onde explicação e resultado sejam obtidos simultaneamente, sejam via junção de dados ou sendo intrínseco a arquitetura do modelo. [16]

- ***Post-Modelling Explainability*:** Também chamada de *Post-hoc Explainability* e onde a maioria dos trabalhos disponíveis baseados em XAI tenta focar, esta fase se

dedica a explicar modelos desenvolvidos anteriormente, ou modelos que não são considerados como transparentes. Pode-se explicar um modelo por métodos como árvores de decisão e estimativas como valores de Shapley e propagação inversa. [17]

Entretanto, estas 3 fases não necessariamente são executadas de maneira sequencial: Por ser realizada no conjunto de dados e não no modelo, a fase de *Pre-modelling Explainability* pode ser opcional, e a aplicação das fases de *Explainable Modelling* e *Post-Modelling Explainability* e seus métodos dependem exclusivamente se o modelo é classificado como transparente ou não, e tal critério é determinado pela adoção de uma (ou mais) das seguintes características [7]:

- **Simulabilidade:** Denota a capacidade de um modelo de ser simulado ou pensado estritamente por um humano. A complexidade assume um lugar dominante nesta classe: Sistemas com uma grande quantidade de regras, por mais simples que essas sejam, já não podem ser classificados como tal. O modelo que se adequa a essa característica precisa ser autocontido o suficiente para que um ser humano pense e raciocine sobre ele como um todo.

- **Decomponibilidade:** Também considerado como inteligibilidade, significa a capacidade de explicar cada uma das partes de um modelo. Se o modelo não for simples o suficiente, ele precisa ser divisível em várias pequenas partes e cada parte do modelo deve ser compreensível por um ser humano, sem a necessidade de ferramentas adicionais.

- **Transparência do algoritmo:** Trata-se da habilidade do usuário de entender o processo seguido pelo modelo para produzir qualquer saída dada a partir de seus dados de entrada. Colocando de outra forma, um modelo linear é considerado transparente porque sua superfície de erro pode ser entendida e fundamentada, permitindo ao usuário entender como o modelo irá agir em cada situação que pode enfrentar.

2.2 Métricas de Fairness

Do ponto de vista social, explicabilidade pode ser considerada como a capacidade de alcançar e garantir a equidade nos modelos de ML, sugerindo uma visualização clara das relações que afetam um resultado e permitindo uma análise da justiça ou eticidade do modelo em questão. Da mesma forma, um objetivo relacionado do XAI é destacar o viés nos dados aos quais um modelo foi exposto. O suporte de algoritmos e modelos está crescendo rapidamente em campos que envolvem vidas humanas, portanto, a explicabilidade deve ser considerada como uma ponte para evitar o uso injusto ou antiético dos resultados do algoritmo [7].

É possível descrever o conceito de *Fairness* no contexto de aprendizagem supervisionada, onde um modelo f pode prever um conjunto de resultados y a partir de um conjunto de *features* x, evitando discriminação injusta em relação a um atributo protegido a (por exemplo, sexo ou raça). É permitido, mas não exigido, que a seja um componente de x [9]. Em outras palavras, um modelo de ML considerado justo é aquele onde a correlação de seu resultado é baixa em relação a dados de entrada considerados como sensíveis a discriminações.

Para medir se um algoritmo é considerado justo ou não, é possível determinar métricas para atribuir um valor. Em um classificador binário, por exemplo, onde dados com resultados reais alimentam o algoritmo para que o mesmo possa estabelecer previsões com outros dados, podemos considerar certas métricas como primitivas [23]:

- **Verdadeiro positivo (TP):** Um caso em que os resultados previstos e reais estão ambos na classe positiva.

- **Falso positivo (FP):** Um caso previsto para estar na classe positiva quando o resultado real pertence à classe negativa.
- **Falso negativo (FN):** Um caso previsto para estar na classe negativa quando o resultado real pertence à classe positiva.
- **Verdadeiro negativo (TN):** Um caso em que os resultados previstos e reais estão ambos na classe negativa.
- **Precisão, ou valor preditivo positivo (PPV):** É a fração de casos positivos previstos corretamente para estar na classe positiva de todos os casos positivos previstos no modelo. Ou seja, $PPV = \frac{TP}{TP+FP}$
- **Taxa de descoberta falsa (FDR):** É a fração de casos negativos previstos incorretamente como estando na classe positiva de todos os casos positivos previstos no modelo. Ou seja, $FDR = \frac{FP}{TP+FP}$
- **Taxa de falsa omissão (FOR):** É a fração de casos positivos previstos incorretamente como estando na classe negativa de todos os casos negativos previstos no modelo. Ou seja, $FOR = \frac{FN}{TN+FN}$
- **Valor preditivo negativo (NPV):** É a fração de casos negativos previstos corretamente para estar na classe negativa de todos os casos negativos previstos no modelo. Ou seja, $NPV = \frac{TN}{TN+FN}$
- **Sensibilidade ou *recall*, ou taxa positiva verdadeira (TPR):** É a fração de casos positivos previstos corretamente para estar na classe positiva de todos os casos positivos reais. Ou seja, $TPR = \frac{TP}{TP+FN}$
- **Taxa de falsos positivos (FPR):** É a fração de casos negativos previstos incorretamente como estando na classe positiva de todos os casos negativos reais. Ou seja, $FPR = \frac{FP}{FP+TN}$
- **Taxa de falsos negativos (FNR):** É a fração de casos positivos previstos incorretamente como estando na classe negativa de todos os casos positivos reais. Ou seja, $FNR = \frac{FN}{TP+FN}$
- **Taxa negativa verdadeira (TNR):** É a fração de casos negativos previstos corretamente para estar na classe negativa de todos os casos negativos reais. Ou seja, $TNR = \frac{TP}{FP+TN}$

Entretanto, tais métricas não agem sozinhas, pois as mesmas por si só já eram utilizadas para aferir a qualidade de modelos desenvolvidos no esquema *black box*. Elas estabelecem uma diretriz inicial, mas dependem de uma análise dos dados para serem usadas de maneira correta e determinar uma medição acurada. Por exemplo, é possível isolar o conjunto de dados por dados como sexo e/ou cor de pele em sub conjuntos, calculá-las criando novas métricas e, por fim, comparar se as métricas possuem os mesmos valores ou estabelecem uma margem segura de tolerância. Dessa forma, é possível afirmar a determinada *Fairness* de um modelo.

2.3 Arquitetura e Engenharia de Software para sistemas de IA

Quando se fala de Arquitetura e Engenharia de Software, se fala da definição dos componentes de software, suas propriedades externas, e seus relacionamentos com outros softwares para fazer com que um sistema seja documentável, reusável e testável. A preocupação está em como um sistema deve ser organizado e com a estrutura geral desse sistema. Quais os requisitos, que ferramentas e equipamentos utilizar, qual o processo de preparação dos dados, o que é possível componentizar e testar são algumas das perguntas que podem envolver uma aplicação baseada em dados e IA.

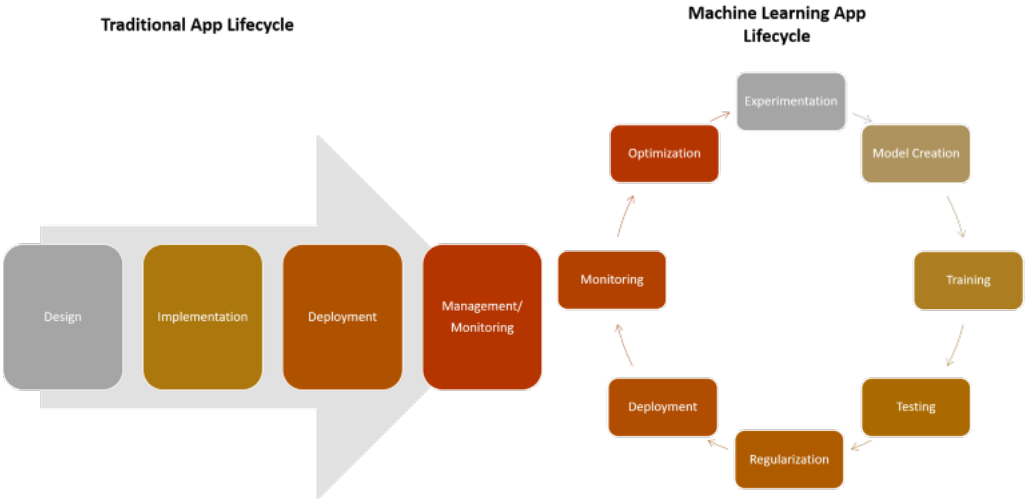

Figura 2: Diferenças entre ciclos de vida de aplicações de software e aplicações de ML. [21]

Em primeiro lugar, uma solução baseada em IA possui um ciclo de vida diferente de uma aplicação tradicional. Conforme ilustrado na Figura 2, sua implementação envolve etapas como criação do modelo e treinamento, e sua fase de manutenção envolve otimizações devido aos novos volumes de dados que chegam em um determinado período de tempo, fazendo com que o modelo esteja sempre em constante mudança. Dessa forma, metodologias ágeis e ferramentas de integração contínua podem não cumprir seus objetivos de maneira ótima [21].

Além disso, o ciclo de vida da aplicação é apenas um dos pontos a se considerar, e as vezes o problema de uma aplicação pode estar em como o dado é obtido e disponibilizado. Um exemplo de arquitetura que generaliza todo o processo, desde a necessidade de negócio até o deploy do modelo de IA, é a IBM Analytics and AI Reference Architecture [2], ilustrada na Figura 3. Nela, são definidos os seguintes requisitos não-funcionais: Performance, estabilidade, segurança, escalabilidade, manutenibilidade e regulamentações de privacidade/*compliance*, e pode ser classificada em 4 grupos principais envolvendo diversos tipos de processos e componentes:

- **Coleta:** Relaciona os processos de coleta, armazenamento e transformação de diversas fontes de dados, estruturadas ou não estruturadas, para determinados repositórios (*Data Lakes*).
- **Organização:** Relaciona os processos de organização e estruturação dos dados nos presentes nos *Data Lakes* e necessários para o uso das aplicações que envolvem a análise dos dados. Dependendo do uso pode-se aplicar processos de Governança.
- **Análise:** Relaciona os processos para desenvolvimento de aplicações de IA e relatórios após a organização dos dados para tomadas de decisão e uso de aplicações externas.

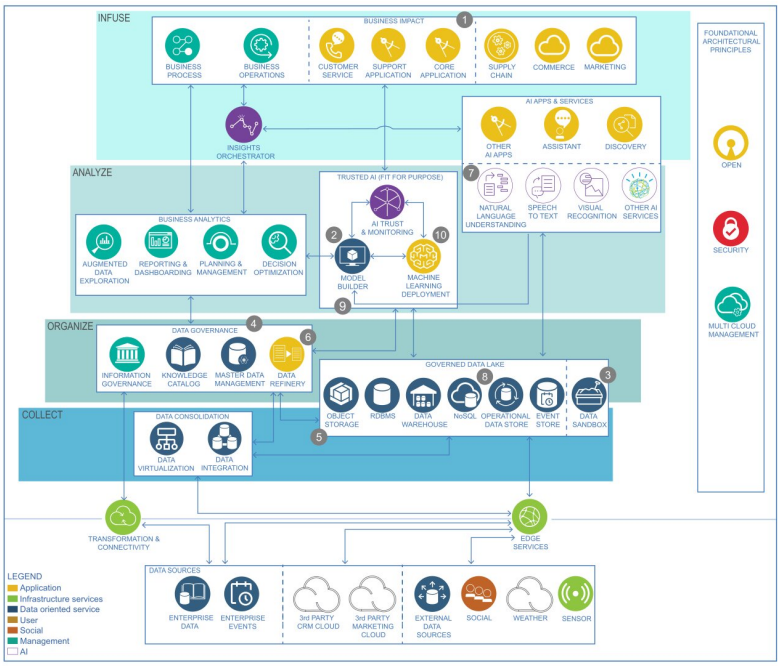

Figura 3: IBM Analytics and AI Reference Architecture.

- **Infusão:** Relaciona os processos de disponibilização desses dados e conhecimento obtidos na fase de análise para aplicações externas.

3 Trabalhos Relacionados

Os principais trabalhos encontrados envolvendo *AI Explainability*, métricas de *Fairness* e sua arquitetura vêm do time de engenharia do LinkedIn. Geyik et al. [12] falam sobre as métricas utilizadas e os algoritmos implementados sobre ranqueamento dos dados. Kenthapadi et al. [14] fala sobre como o conceito de *Fairness* pode ser implementado em uma aplicação de IA, e mostra a arquitetura sobre ela, assim como outros estudos de caso [15].

No ramo de IA, muitos dos trabalhos desenvolvidos se referem a novas técnicas e modelos e poucos se referem realmente a arquitetura de um sistema. Entretanto, como os conceitos de *AI Explainability* e métricas de *Fairness* agem mais como extensões para construir modelos mais resilientes e não como novos paradigmas de modelagem, também podemos usar como referência arquiteturas de software já existentes e adaptá-las para já considerar estes conceitos desde sua base.

Bailis et al. [8] propuseram em Stanford o projeto DAWN, que propõe uma série de ferramentas, *frameworks* e uma arquitetura de referência para otimizar workflows de ML. Embora seja focada mais no processo de ML do que em XAI, uma de suas direções foca em elaborar interfaces para especialistas fora da área de ML e explicação de resultados e interpretações para humanos.

Modi et al. [20] propuseram no Google o TFX, uma plataforma de ML para uso geral e baseada no TensorFlow, *framework* do Google especializado em *Deep Learning*. Ao integrar os componentes em uma plataforma, foram capazes de padronizar os componentes, simplificar a configuração da plataforma e reduzir o tempo de produção de meses para semanas, ao mesmo tempo que era fornecida estabilidade. Em um de seus trabalhos futuros, foi mencionada a necessidade de explicação das decisões e ações realizadas pelos modelos, conforme o uso de ML fosse se intensificando. O Google atualmente possui uma ferramenta para *AI Explainability* em sua plataforma de computação em nuvem [1] possuindo métodos já imple-

mentados e disponibiliza explicações do modelo via metadados, mas nada foi encontrado em como isso era relacionado na arquitetura de um sistema real.

Uma outra arquitetura que é possível considerar e se basear é a FBFlow do Facebook [11]. Embora seja mais específica para a implementação e não possua qualquer menção a necessidades para explicação, o uso de DAGs (*Direct Acyclic Graphs*) para a execução de um determinado processo ou *workflow* é uma alternativa simples de ser implementada e interpretada pelos desenvolvedores, onde é possível conectar o método de ML juntamente com o cálculo de métricas e métodos de XAI para serem automatizados e organizados de forma simultânea.

Referências	**Descrição**	Arquitetura		Conceitos		
		Arq. Referência	**Arq. Sistema**	**XAI**	**ML**	**Fairness**
Geyik et al., 2019	Métricas utilizadas Ranqueamento dos dados		✓		✓	✓
Kenthapadi et al., 2019	Arquitetura/implementação de métricas de Fairness		✓		✓	✓
Kenthapadi et al., 2020	Desafios/lições sobre Explainable AI		✓	✓	✓	
IBM Analytics and AI Reference Architecture	Arquitetura de referência para Dados e ML	✓			✓	
Projeto DAWN - Stanford (Ballis et al., 2017)	Ferramentas/Frameworks Arquitetura de referência para ML	✓		Trab. Futuro	✓	
TFX - Google (Modi et al., 2017)	Componentização/Padronização Simplificação na implementação		✓	Trab. Futuro	✓	
FBFlow - Facebook (Dunn, 2016)	Uso de DAGs para determinar *workflow* Execução/interpretação de processos de ML		✓		✓	

4 Resultados Esperados e Desafios

Espera-se obter como resultado desse projeto de mestrado uma proposta de arquitetura de software que acople os conceitos de *Explainable AI* e métricas de *Fairness* em uma solução de IA, com uma elaboração de caso de uso feito de acordo com as diretrizes obtidas neste projeto. Espera-se também que tal arquitetura se adeque independentemente de como os dados serão armazenados e processados.

O foco desse projeto não está necessariamente em obter resultados melhores em um modelo, e sim que a melhora dos resultados seja consequência de sua explicabilidade e de uma manutenção menos custosa. Espera-se também a elaboração e publicação de artigos em periódicos e eventos relacionados, relatando as experiências com o projeto e os resultados obtidos.

Por ser um campo bastante amplo e em constante evolução, este projeto terá limitações de escopo e voltará apenas para o aprendizado supervisionado, que é mais simples de ser explicado por já conhecermos suas entradas e saídas e por muitas soluções de IA serem treinadas com essa abordagem [10]. Entretanto, é esperado que este projeto contribua com diretrizes que possam se aplicar também a modelos não supervisionados e de aprendizado por reforço.

Embora haja várias ferramentas e algoritmos baseados em *Explainable AI*, as mesmas estão dispersas e há pouco material destinado a organizar estes tipos de ferramentas. Pretende-se solucionar isso com a elaboração de *Checklists* e um caso de uso para estabelecer exemplos de desenvolvimento.

Embora as soluções de *Explainability AI* ajudam a mesma a ser testável devido ao rastreamento do que está sendo notado para elaborar o resultado, é possível colocar como desafio que pontos estas soluções ajudam em específico. Como solução, é preciso realizar um estudo, assim como inclusão em documentação e exemplos no caso de uso.

Outro desafio está em como conectar os diferentes métodos utilizados com o método de ML e o conjunto de dados adequado. Para isso, componentes baseados na arquitetura MAPE-K [3] podem ser úteis, podendo estabelecer uma base de conhecimento e estabelecer relações entre os métodos de forma adequada.

5 Cronograma

Conforme o programa de mestrado, o aluno realizará as disciplinas nos dois primeiros semestres, e durante o terceiro semestre participará do Programa de Estágio a Docência (PED) da Unicamp. Após a defesa do Exame de Qualificação de Mestrado (EQM), o aluno continuará a elaborar um estudo de caso e começará a redigir sua Dissertação. Ao final do segundo ano de projeto, será realizada a Dissertação de Mestrado do aluno.

As etapas que envolvem este projeto de Mestrado começarão no segundo semestre devido a uma desistência de orientação anterior do aluno, sendo possível ter uma prorrogação para um quinto semestre para a finalização do curso.

A Tabela abaixo mostra a estrutura do cronograma do projeto.

	2020	2021		2022
Atividade	**Semestre 1**	**Semestre 2**	**Semestre 3**	**Semestre 4**
Realização de disciplinas do IC	✓	✓		
Estudo e catálogo de artigos		•		
Revisão bibliográfica		•	•	•
Elaboração de caso de uso			•	•
Qualificação		•		
Escrita da dissertação			•	•
Realização do PED			•	
Publicação de artigo				•
Defesa				•

Referências

[1] Google - AI Explanations Whitepaper. URL https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf.

[2] IBM - Analytics and AI architecture. URL https://www.ibm.com/cloud/architecture/architectures/aiAnalyticsArchitecture/reference-architecture/.

[3] An architectural blueprint for autonomic computing. Technical report, IBM, 2005. URL https://www-03.ibm.com/autonomic/pdfs/AC%20Blueprint%20White%20Paper%20V7.pdf.

[4] Machine learning: things are getting intense, 2018. URL https://www2.deloitte.com/content/dam/Deloitte/global/Images/infographics/technologymediatelecommunications/gx-deloitte-tmt-2018-intense-machine-learning-report.pdf.

[5] Brasil se destaca com 42% das iniciativas de IA na América Latina, em 2020, 2021. URL https://cio.com.br/tendencias/brasil-se-destaca-com-42-das-iniciativas-de-ia-na-america-latina-em-2020/.

[6] Proteção de Dados - LGPD, 2021. URL https://www.gov.br/defesa/pt-br/acesso-a-informacao/lei-geral-de-protecao-de-dados-pessoais-lgpd.

[7] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[8] Peter Bailis, Kunle Olukoton, Christopher Re, and Matei Zaharia. Infrastructure for usable machine learning: The stanford dawn project. 2017.

[9] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. Explainability for fair machine learning, 2021.

[10] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, 2018.

[11] Jeffrey Dunn. Introducing FBLearner Flow: Facebook's AI backbone. URL https://engineering.fb.com/2016/05/09/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/.

[12] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-Aware Ranking in Search and Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining*, pages 2221–2231, 2019.

[13] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda B. Viégas, and Michael Terry. XRAI: better attributions through regions. In *IEEE/CVF International Conference on Computer Vision*, pages 4947–4956, 2019.

[14] Krishnaram Kenthapadi, Stuart Ambler, Ahsan Chudhary, Mark Dietz, Sahin Cem Geyik, Ian Koeppe, Varun Mithal, Guillaume Saint-Jacques, Amir Sepehri, Thanh Tran, and Sriram Vasudevan. Fairness, privacy, and transparency by design in ai/ml systems, 2019. URL https://engineering.linkedin.com/blog/2019/fairness-privacy-transparency-by-design.

[15] Krishnaram Kenthapadi, Sahin Cem Geyik, Varun Mithal, Krishna Gade, and Ankur Taly. Explainable AI in Industry: Practical Challenges and Lessons Learned, 2020. URL https://sites.google.com/view/www20-explainable-ai-tutorial.

[16] Bahador Khalegi. The how of explainable ai: Explainable modelling, 2019. URL https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed.

[17] Bahador Khalegi. The how of explainable ai: Post-modelling explainability, 2019. URL https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f.

[18] Bahador Khalegi. The how of explainable ai: Pre-modelling explainability, 2019. URL https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4.

[19] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. 2013.

[20] Akshay Naresh Modi, Chiu Yuen Koo, Chuan Yu Foo, Clemens Mewald, Denis M. Baylor, Eric Breck, Heng-Tze Cheng, Jarek Wilkiewicz, Levent Koc, Lukasz Lew, Martin A. Zinkevich, Martin Wicke, Mustafa Ispir, Neoklis Polyzotis, Noah Fiedel, Salem Elie Haykal, Steven Whang, Sudip Roy, Sukriti Ramesh, Vihan Jain, Xin Zhang, and Zakaria Haque. Tfx: A tensorflow-based production-scale machine learning platform. In *KDD 2017*, 2017.

[21] Jesus Rodriguez. Machine learning reference architectures from google, facebook, uber, databricks and others, 2020. URL https://medium.com/dataseries/machine-learning-reference-architectures-from-google-facebook-uber-databricks-a

[22] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 2017.

[23] Sahil Verma and Julia Rubin. Fairness definitions explained. In *ACM/IEEE International Workshop on Software Fairness*, pages 1–7, 2018.