**Adriano Alonso Veloso**

e-mail: adrianov@dcc.ufmg.br

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

# HarmonyML: Aprendizado de modelos alinhados aos valores humanos e às restrições de negócio

Projeto apresentado como requisito à renovação da Bolsa de Produtividade em Pesquisa do CNPq.

## Projeto Anterior (Agosto 2020 – Julho 2023)

O proponente tem bolsa de produtividade em curso, e portanto a seguir são descritas as metas referentes à proposta anterior, bem como os resultados obtidos.

**Metas do Projeto Anterior:** (i) projetar, implementar e validar novos algoritmos que façam uso de restrições de explicação durante o processo de aprendizado de modelos preditivos; (ii) avaliar a efetividade prática dos algoritmos desenvolvidos em tarefas relevantes e desafiadoras; (iii) formar 3 doutores, 8 mestres e 8 alunos de iniciação científica; (iv) publicar 6 artigos em periódicos, 10 artigos em conferências internacionais e 6 artigos em conferências nacionais; e (v) publicar e prover acesso ao estado-da-arte à academia e empresas para que essas possam se beneficiar de oportunidades abertas pelo projeto proposto.

**Resultados Alcançados:** (i), (ii) e (v) métodos e algoritmos de aprendizado profundo foram projetados, implementados, validados, e disponibilizados. Um novo modelo de transferência de tecnologia foi criado, culminando com a criação do Laboratório de Inteligência Artificial (LIA), onde pelo menos 30 projetos são realizados em parceria com diferentes empresas e universidades; (iii) 6 alunos (Geanderson Esteves, Tiago Amador, Anderson Bessa, Amir Jallilifard, Gianlucca Zuin, e Daniella Araújo) tiveram suas teses aprovadas [1, 4, 5, 7, 11, 17], e um aluno (Ismael Santana) passou pela qualificação de tese [16] e será doutor até 2024. Formamos 10 mestres (Francisco Galuppo [8], Silvia Guerra [10], Felipe Glicério [9], Victor Rodrigues [15], Guilherme Mendes [14], Camila Kolling [12], Eduardo Nigri [13], João Miranda [2], Lucas Aquino [3], André Correia [6]), e 10 alunos de iniciação científica (Luan Borges, Júlia Manuela, Alefi Santos, Matheus Freitas, Daniel Campos, Augusto Maillo, Luis Felipe Ramos, Lorenzo Carneiro, Bruno dos Santos Lopes, Felipe Santos); (iv) publicamos 16 artigos em periódicos [33–41, 43–45, 47–50], 13 artigos em conferências internacionais [18–22, 24, 27–30, 44, 63, 67] (outros 11 artigos estão submetidos e em processo de avaliação por pares com desfecho nos próximos meses).

## Atuação no Período – 2020 a 2023

De acordo com a Chamada CNPq No 09/2023 – Bolsas de Produtividade em Pesquisa, ANEXO I – Critérios dos Comitês Assessores, o item "Critérios Específicos" do documento CC – Ciência da Computação, menciona: "*De forma complementar, têm sido levados em consideração outros indicadores objetivos tais como orientações concluídas, total de recursos obtidos em projetos de pesquisa, prêmios e distinções recebidas e participação em comitês científicos. Nas atividades de orientação, alguns aspectos analisados são: quantos alunos de mestrado/doutorado concluíram suas dissertações/teses sob sua orientação no período relevante para o julgamento. Que trabalhos associados a essas orientações foram publicados ou submetidos para publicação em periódicos e/ou eventos nacionais e/ou internacionais. Quantas orientações de mestrado e doutorado estão em andamento. Que tipos de cursos de pós-graduação relacionados à sua pesquisa o proponente tem lecionado, e com que regularidade. Organização de eventos. Criação de startups e interações com empresas. Depósito de patentes. Premiações.*"

Assim sendo, apresentamos a seguir, de forma mais detalhada, indicadores objetivos a respeito da atuação acadêmica do proponente no período relevante para o julgamento (2020 a 2023).

**Orientações Concluídas:** Foram defendidas 6 teses de doutorado e 10 dissertações de mestrado, detalhadas a seguir.

1. Daniella Araújo – DAS: synthetic generation for medical data augmentation and smoothing, 08/08/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Agma Juci Machado Traina (ICMC-USP), Karina Braga (UFMG), Gabriel Coutinho (UFMG), Nivio Ziviani (UFMG). Artigos publicados: [34–36, 41, 45].

2. Amir Jallilifard – Modeling pharmacological effects through multi-graph and multi-relation graph embedding, 27/02/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Leandro Balby (UFCG), Duncan Ruiz (PUC-RS), Nivio Ziviani (UFMG), Marcos dos Santos (UFMG), Renato Vimiero (UFMG). Artigos publicados: [19, 21].

3. Geanderson Esteves dos Santos – Understanding software defects with Machine Learning, 13/02/2023. Banca examinadora: Eduardo Figueiredo (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Ivan Machado (UFBA), Valter de Camargo (UFSCAR), Marco Túlio Valente (UFMG), Wagner Meira Jr. (UFMG). Artigos publicados: [31, 32], [20], [39].

4. Gianlucca Zuin – Ensemble learning through Rashomon sets, 05/01/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Rafael Bordini (PUC-RS), Ram Rajagopal (Stanford University), Paulo Orenstein (IMPA), Nivio Ziviani (UFMG), Wagner Meira Jr (UFMG). Artigos publicados: [28–30], [48–50].

5. Tiago Amador – Regularização de modelos para predição precoce: Um estudo sobre a predição de complicações na UTI, 30/08/2022. Banca examinadora: Adriano Veloso (UFMG, orientador), Renato Vimiero (UFMG), Saulo Saturnino (UFMG), Soraia Musse (PUC-RS), Wagner Meira Jr. (UFMG), Leandro Balby (UFCG). Artigos publicados: [33].

6. Anderson Bessa – Ensemble learning by diversifying explanations: Predicting the evolution of pain relief, 20/12/2021. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG, coorientador), Wagner Meira Jr. (UFMG), Leandro Balby (UFCG), Marco Cristo (UFAM), Daniel Ciampi (Aalborg University). Artigos publicados: [38].

7. Francisco Galuppo – Stochastic Neural Dynamic Programming, 08/08/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Wagner Meira Jr. (UFMG), Ram Rajagopal (Stanford University). Artigo submetido e em avaliação por pares.

8. Lucas Aquino – Assessment of Agricultural Production Capacity Through Remote Sensing and Machine Learning, 08/08/2023. Banca examinadora: Nivio Ziviani (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Heitor Ramos Filho (UFMG), Edleno Moura (UFAM). Artigo submetido e em avaliação por pares.

9. André Correia – Truth or Utility: Transductive regularizer for feature selection, 01/03/202. Banca examinadora: Nivio Ziviani (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Heitor Soares Ramos Filho (UFMG), Anderson Soares (UFG). Artigos publicados: [67], [42].

10. Felipe Glicério – Detecção automática de glaucoma primário usando algoritmos de aprendizado de profundo, 17/02/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), Sebastião Cronemberger (UFMG). Artigo submetido e em avaliação por pares.

11. Silvia Guerra – Contextual NLP Explanations for Language Biomarker Research: Identification of Schizophrenia Traits on Social Media Posts using Multilevel Part of Speech Feature, 16/02/2023. Banca examinadora: Adriano Veloso (UFMG, orientador), Cilene Rodrigues (PUC-RJ), Nivio Ziviani (UFMG), Anderson Bessa (UFGD). Artigo submetido e em avaliação por pares.

12. Camila Kolling – Mitigating bias in facial analysis systems by incorporating label diversity, 21/07/2022. Banca examinadora: Soraia Musse (PUC-RS, orientadora), Adriano Veloso (UFMG, coorientador), (UFMG), Virgilio Almeida (UFMG), Rafael Bordini (PUC-RS). Artigo submetido e em avaliação por pares.

13. João Miranda – Pronomes em esquizofrenia: análise de textos escritos no contexto de mídia social, 20/04/2022. Banca examinadora: Cilene Rodrigues (PUC-RJ, orientadora), Adriano Veloso (UFMG, coorientador), Rafael Moreno (UFCSPA), Eduardo Kenedy Nunes (UFF). Artigo submetido e em avaliação por pares.

14. Guilherme Mendes – A deep learning model for automatic recognition of pain facial expressions on human fetuses, 24/03/2022. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), Erickson Nascimento (UFMG), George Medeiros (UFMG), Lisandra Stein (USP). Artigo submetido e em avaliação por pares.

15. Victor Rodrigues – Uma abordagem baseada em aprendizado de máquina para a modelagem química do aço inoxidável duplex resistente à formação de lascas de aquecimento, 14/09/2021. Banca examinadora: Adriano Veloso (UFMG, orientador), Heitor Soares Raos Filho (UFMG), Mauricio Marengoni (SENAC-SP). Artigos publicados: [29].

16. Eduardo Nigri – Visual Explanations of Convolutional Neural Networks for MRI Classification of Alzheimer's Disease, 16/11/2020. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), Paulo Caramelli (UFMG), Jefersson dos Santos (UFMG), Rodrigo Barros (PUC-RS). Artigos publicados: [63].

**Orientações em Andamento:** Há 3 orientações de doutorado e 11 orientações de mestrado em curso, detalhadas a seguir.

1. Ismael Santana, doutorado (orientador), defesa prevista para março 2024. Artigos publicados: [44], [46].

2. Roberta Viola, doutorado (orientador), defesa prevista para março de 2026. Artigos publicados: [22, 27].
3. Guilherme Mendes, doutorado (orientador), defesa prevista para março de 2026.
4. Ramon Costa, mestrado (orientador), defesa prevista para março de 2025.
5. Clarissa Lima, mestrado (orientador), defesa prevista para março de 2025.
6. Luiz Henrique de Melo, mestrado (orientador), defesa prevista para março de 2024.
7. Guilherme Drummond, mestrado (orientador), defesa prevista para março de 2024. Artigos publicados: [27].
8. Pedro Martins, mestrado (orientador), defesa prevista para março de 2024.
9. Mychell Laurindo, mestrado (orientador), defesa prevista para setembro de 2023. Artigos publicados: [29].
10. Marco Antônio Tavares, mestrado (orientador), defesa prevista para março de 2024.
11. Douglas Pontes, mestrado (orientador), defesa prevista para março de 2024.
12. Luiz Felipe Viana, mestrado (orientador), defesa prevista para setembro de 2023.
13. Djim Martins, mestrado (orientador), defesa prevista para março de 2024.
14. Késia Dias, mestrado (orientador), defesa prevista para outubro de 2023.

**Supervisões de Pós-Doutorado:** Foram concluídas três supervisões de pós-doutorado e outras duas supervisões estão em curso:

1. Solange Mata Machado, Impacto do mindset na geração de projetos de inteligência artificial. Início: 2017/02. Encerramento: 2019/02 (encerrado)
2. Raquel Fabreti de Oliveira, Aplicação de metodologias da inteligência artificial para predizer fatores de risco e sobrevida no transplante renal. Início: 2020/01 (encerrado)
3. Rafael Moreno, Inteligência Artificial em Psiquiatria, Início: 2021/10 (encerrado)
4. Fabiana Menezes, Análise preditiva aplicada a sistemas normativos complexos: IA para detecção de riscos ao direito à alimentação. Início: 2021/09 (em curso).
5. Saulo Saturnino, Aprendizado de máquina como ferramenta de predição, otimização de resultados, e redução de carga de trabalho no cuidado de pacientes críticos. Início: 2022/09 (em curso).

**Recursos Obtidos/Projetos de Pesquisa:** No período, o proponente atuou/atua como coordenador de cinco projetos de pesquisa, que são detalhados a seguir.

1. Laboratório de Inteligência Artificial − Projeto de Parceria de Pesquisa com a empresa Kunumi. Valor: R$1.650.000,00.

2. Predição de Defeitos durante a Produção de Aço através de Aprendizado Causal-Explicativo − Projeto de Parceria de Pesquisa com a empresa ArcelorMittal. Valor: R$140.727,27.

3. Modelos de Aprendizado de Máquina para Identificação de Falhas durante o Processo de Desodorização de Óleos − Projeto de Parceria de Pesquisa com a empresa ST-One. Valor: R$116.470,59.

4. Modelos de Inteligência de Dados para o Mercado de Materiais de Construção − Projeto de Parceria de Pesquisa com a empresa Dexco. Valor: R$240.000,00.

5. Modelos de Inteligência Artificial para People Analytics − Projeto de Parceria de Pesquisa com a empresa Dexco. Valor: R$148.235,29.

6. Participação como Membro Especialista de Comitê Consultivo da Dexco − Projeto de Consultoria com a empresa Dexco. Valor: R$360.000,00.

7. Aprendizado de máquina como ferramenta de predição, otimização de resultados, e redução de carga de trabalho no cuidado de pacientes críticos − Centro de Inovação em Inteligência Artificial para a Saúde. Valor: R$80.000,00.

8. SafeML: Aprendizado de modelos preditivos seguros com restrições quanto ao mecanismo de predição, CNPq Bolsa de Produtividade em Pesquisa, nível 2. Valor: R$39.600,00.

No período, o proponente também participa/participou de outros cinco projetos, que são detalhados a seguir:

1. Aprendizado de Representações de Pacientes para Previsão de Tempo de Internação − Centro de Inovação em Inteligência Artificial para a Saúde. Valor: R$80.000,00.

2. CIIA-Saúde: Centro de Inovação em Inteligência Artificial para Saúde − Fapemig. Valor: R$5.000.000,00.

3. Capacidades Analíticas do Ministério Público de Minas Gerais − Ministério Público de Minas Gerais. Valor: R$4.074.556,21

4. Inteligência artificial aplicada à exploração de petróleo na camada pré-sal − Petrobras. Valor: R$12.112.945,00

5. INCT-Cyber: Instituto Nacional de CIência e Tecnologia para uma Sociedade Massivamente Conectada − MCT/CNPq.

Como detalhado, ao longo dos últimos três anos o proponente participou de projetos de pesquisa que totalizam R$23.802.533,00, dos quais R$2.535.032,00 são relativos a projetos coordenados pelo proponente.

**Atuação na Comunidade Nacional:** O proponente tem atividades relevantes na comunidade científica da Computação no país. Foi membro afiliado da Acadêmica Brasileira de Ciências (Engenharias) e atuou ativamente como membro do comitê de programa de conferências nacionais e como revisor de periódicos nacionais:

- Revisor para JIDM (Journal of Information and Data Management),
- Revisor para JBCS (Journal of the Brazilian Computer Society),
- Revisor para REIC (Revista Eletrônica de Iniciação Científica da SBC),
- Revisor para RITA (Revista de Informática Teórica e Aplicada).
- Membro do comitê de programa da BRASNAM (Brazilian Workshop on Social Network Analysis and Mining), do CTD (Concurso de Teses e Dissertações da Sociedade Brasileira de Computação), do Bracis (Brazilian Conference on Intelligent Systems), da WebMedia (Simpósio Brasileiro de Sistemas Multimídia e Web), do CTIC (Concurso Nacional de Trabalhos de Iniciação Científica da Sociedade Brasileira de Computação), do KDMile (Brazilian Symposium on Knowledge Discovery, Mining and Learning), e do SIBGRAPI (Brazilian Conference on Graphics, Patterns and Images).
- Organizador/Moderador do Painel "From Data into Energy: AI Reshaping Offshore" da OTC 2023 (Offshore Technology Conference).

O proponente também tem atuado como revisor de projeto de fomento para as agências: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Fundação de Amparo a Pesquisa do Estado do Amazonas (FAPEAM), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Atuação na Comunidade Internacional:** O proponente tem trabalhado como revisor dos mais importantes periódicos das áreas de Aprendizado de Máquina e Processamento de Linguagem Natural:

- Data & Knowledge Engineering − Elsevier,
- Distributed and Parallel Databases − Springer,
- Information Sciences − Elsevier,
- Information Systems − Elsevier,
- Information Processing & Management − Elsevier,
- The VLDB Journal − VLDB Endownment,
- Transactions on Knowledge and Data Engineering − IEEE,
- Transactions on Parallel and Distributed Systems − IEEE,

- Data Mining and Knowledge Discovery – Springer,
- Journal of Autonomous Agents and Multi-Agent Systems – Springer,
- Transactions on Intelligent Systems and Technology – ACM,
- Transactions on Knowledge Discovery from Data – ACM,
- Knowledge Engineering Review – Cambridge,
- Security and Communication Networks – Wiley

O proponente tem sido membro do comitê de programa dos seguintes eventos científicos internacionais[1]:

- AAAI Conference on Artificial Intelligence, desde 2018,
- Annual Meeting of the Association for Computational Linguistics, desde 2017,
- IEEE Big Data Conference, desde 2016,
- Internation Joint Conference on Artificial Intelligence, desde 2015,
- ACM CIKM Conference on Information and Knowledge Management, desde 2014,
- SIAM International Conference on Data Mining, desde 2014,
- AAAI International Conference on Web and Social Media, desde 2014,
- International Joint Conference on Natural Language Processing, desde 2014,
- ACM Web Science Conference, desde 2013,
- IEEE International Conference on Data Mining, desde 2012,
- String Processing and Information Retrieval Conference, desde 2012,
- World Wide Web Conference, desde 2012,
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining, desde 2012,
- ACM SIGIR Conference on Research and Development in Information Retrieval, desde 2011.
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, desde 2010,

[1]Em algumas edições o proponente eventualmente precisou recusar o convite por motivo de indisponibilidade.

Também apresentamos na Tabela 1 o sumário de indicadores de produtividade do proponente nos últimos 5 anos e nos últimos 10 anos. Informações mais detalhadas e atualizadas podem ser encontradas no currículo Lattes do proponente[2]. Observa-se uma média de aproximadamente 2,2 publicações por aluno de pós-graduação. Além disso, cerca de 80% das publicações do proponente estão no extrato superior, e como pode ser visto na Tabela 2, nos últimos anos o proponente vem focado em publicações no extrato A. O proponente publica aproximadamente um artigo em periódico a cada três artigos em evento.

Ainda na Tabela 1, ressalta-se a patente entitulada "PROCESSO CENTRADO NO HUMANO PARA ELABORAÇÃO DE MODELOS BASEADOS EM APRENDIZADO DE MÁQUINA E USOS" cujo teor técnico é intimamente relacionado aos objetivos deste projeto de pesquisa. Por fim, ressalta-se a criação de uma spin-off chamada Huna (`https://www.huna-ai.com/`), de ex-alunos do LIA (Laboratório de Inteligência Artificial da UFMG) orientados pelo proponente. A empresa já passou, com êxito, pela fase de aportes e dedica-se a construir modelos diagnósticos alinhados com o entendimento médico sobre diferentes doenças crônicas.

**Cursos de Pós-Graduação:** O proponente leciona regularmente os cursos de Aprendizado de Máquina/Aprendizado Profundo (desde 2010), e de Processamento de Linguagem Natural (desde 2015). A Tabela 3 mostra o número de alunos matriculados nessas disciplinas anualmente.

# Teses e Dissertações Defendidas – 2020 a 2023

[1] Tiago Amador. Regularização de modelos para predição precoce: Um estudo sobre a predição de complicações na UTI (Doutorado), 2022.

[2] Jo ao Victor Miranda. Pronomes em esquizofrenia: análise de textos escritos no contexto de mídia social (Mestrado), 2022.

[3] Lucas Aquino. Assessment of agricultural production capacity through remote sensing and machine learning (Mestrado), 2023.

[4] Daniella Araújo. Das: synthetic generation for medical data augmentation and smoothing (Doutorado), 2023.

[5] Anderson Bessa. Explainability, predictability and counterfactuals: An alternative to all-in-one approach (Doutorado), 2021.

[6] André Correia. Truth or utility: Transductive regularizer for feature selection (Mestrado), 2023.

[7] Geanderson Esteves. Understanding software defects with machine learning (Doutorado), 2023.

[8] Francisco Galuppo. Stochastic neural dynamic programming (Mestrado), 2023.

[2] `http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4762232P7`

| **Indicador** | **5 anos** | **10 anos** |
|---|---|---|
| H-Index (*Google Scholar*) | 21 | 34 |
| Citações (*Google Scholar*) | 2220 | 4256 |
| Livros | 0 | 1 |
| Periódicos | 29 | 40 |
| Artigos em conferências | 61 | 90 |
| **Orientações concluídas** | 36 | 62 |
| Doutorado | 8 | 9 |
| Mestrado | 26 | 40 |
| Iniciação Científica | 28 | 39 |
| **Orientações em andamento** | 20 | – |
| Doutorado | 3 | – |
| Mestrado | 7 | – |
| Iniciação Científica | 10 | – |
| **Prêmios e distinções científicas** | 12 | 20 |
| Nacional | 11 | 18 |
| Internacional | 1 | 2 |
| **Projetos de pesquisa na própria instituição** | 16 | 22 |
| Coordenação | 11 | 12 |
| Participação | 8 | 13 |
| **Projetos de pesquisa de cooperação internacional** | 2 | 4 |
| Coordenação | 1 | 1 |
| Participação | 2 | 4 |
| **Projetos de pesquisa com org. públicas ou privadas** | 10 | 16 |
| Coordenação | 7 | 9 |
| Participação | 10 | 11 |
| **Revisor de periódicos** | 18 | 22 |
| Nacionais | 4 | 5 |
| Internacionais | 14 | 17 |
| **Eventos** | 21 | 23 |
| TPC de conferências nacionais | 8 | 8 |
| TPC de conferências internacionais | 13 | 15 |
| **Patentes** | 1 | 1 |
| **Spin-Offs criadas** | 1 | 1 |

Table 1: Sumário com indicadores de desempenho em pesquisa.

| | **Eventos** | | **Periódicos** | |
|---|---|---|---|---|
| | **5 anos** | **10 anos** | **5 anos** | **10 anos** |
| A1 | 32 | 50 | 11 | 16 |
| A2 | 6 | 9 | 2 | 5 |
| A3 | 2 | 4 | 2 | 2 |

Table 2: Publicações em eventos e periódicos qualificados (A1 − A3).

Table 3: Disciplinas lecionadas por semestre e com tamanho de cada turma.

| Semestre | Nome da disciplina | Matriculados |
|---|---|---|
| 2010.2 | Aprendizado de Máquina (PG) | 35 (grad) + 41 (pós) |
| 2012.1 | Aprendizado de Máquina (PG) | 17 (grad) + 28 (pós) |
| 2013.1 | Aprendizado de Máquina (PG) | 21 (grad) + 36 (pós) |
| 2014.1 | Aprendizado de Máquina (PG) | 27 (grad) + 31 (pós) |
| 2015.1 | Aprendizado de Máquina (PG) | 30 (grad) + 33 (pós) |
| 2015.2 | Processamento de Linguagem Natural (PG) | 7 (grad) + 14 (pós) |
| 2016.1 | Aprendizado de Máquina (PG) | 34 (grad) + 26 (pós) |
| 2016.2 | Processamento de Linguagem Natural (PG) | 5 (grad) + 17 (pós) |
| 2017.1 | Aprendizado de Máquina (PG) | 28 (grad) + 44 (pós) |
| 2017.2 | Processamento de Linguagem Natural (PG) | 16 (grad) + 22 (pós) |
| 2018.1 | Aprendizado de Máquina (PG) | 25 (grad) + 46 (pós) |
| 2018.2 | Processamento de Linguagem Natural (PG) | 19 (grad) + 22 (pós) |
| 2019.1 | Aprendizado de Máquina (PG) | 29 (grad) + 49 (pós) |
| 2019.2 | Processamento de Linguagem Natural (PG) | 16 (grad) + 25 (pós) |
| 2021.1 | Aprendizado de Máquina (PG) | 34 (grad) + 58 (pós) |
| 2021.2 | Processamento de Linguagem Natural (PG) | 19 (grad) + 28 (pós) |
| 2022.1 | Aprendizado de Máquina (PG) | 31 (grad) + 55 (pós) |
| 2023.1 | Aprendizado de Máquina (PG) | 38 (grad) + 62 (pós) |
| 2010.1−2023.1 | | 431 (grad) + 637 (pós) |


[9] Felipe Glicério. Detecção automática de glaucoma primário usando algoritmos de aprendizado de profundo (Mestrado), 2023.

[10] Silvia Guerra. Contextual NLP explanations for language biomarker research: Identification of schizophrenia traits on social media posts using multilevel part of speech feature (Mestrado), 2023.

[11] Amir Jalilifard. Modeling pharmacological effects through multi-graph and multi-relation graph embedding (Doutorado),, 2023.

[12] Camila Kolling. Mitigating bias in facial analysis systems by incorporating label diversity (Mestrado), 2022.

[13] Eduardo Nigri. Visual explanations of convolutional neural networks for MRI classification of alzheimers disease (Mestrado), 2020.

[14] Guilherme Oliveira. A deep learning model for automatic recognition of pain facial expressions on human fetuses (Mestrado), 2023.

[15] Victor Rodrigues. Uma abordagem baseada em aprendizado de máquina para a

modelagem química do aÃ§o inoxidável duplex resistente à formação de lascas de aquecimento (Mestrado), 2021.

[16] Ismael Santana. Model efficacy estimation on out-of-distribution data by causal regularization (Doutorado), 2023.

[17] Gianlucca Zuin. Ensemble learning through rashomon sets (Doutorado), 2023.

## Conferências Internacionais – 2020 a 2023

[18] Lisandra S Bernardes, Mariana A Carvalho, Simone B Harnik, Manoel J Teixeira, Juliana Ottolia, Daniella Castro, Adriano Veloso, Rossana Francisco, Clarice Listik, Ricardo Galhardoni, Valquiria Aparecida da Silva, Larissa I Moreira, Antonio G de Amorim Filho, Ana M Fernandes, and Daniel Ciampi de Andrade. Sorting pain out of salience: assessment of pain facial expressions in the human fetus. *Pain Reports*, 6(1):63–79, 2021.

[19] Dehua Chen, Amir Jalilifard, Adriano Veloso, and Nivio Ziviani. Modeling pharmacological effects with multi-relation unsupervised graph embedding. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–7, 2020.

[20] Geanderson E. dos Santos, Adriano Veloso, and Eduardo Figueiredo. The subtle art of digging for defects: Analyzing features for defect prediction in java projects. In *International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE*, pages 371–378, 2022.

[21] Amir Jalilifard and Adriano Veloso. Drug repurposing opportunities in shapley space. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2022.

[22] Caio Libânio Melo Jerônimo, Cláudio Elízio Calazans Campelo, Leandro Balby Marinho, Allan Sales da Costa Melo, Adriano Veloso, and Roberta Viola. Computing with subjectivity lexicons. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*, pages 3272–3280, 2020.

[23] Eduardo Nigri, Nivio Ziviani, Fabio A. M. Cappabianco, Augusto Antunes, and Adriano Veloso. Explainable deep cnns for mri-based diagnosis of alzheimers disease. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2020.

[24] Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2020.

[25] Julio C. S. Reis, André Correia, Fabricio Murai, Adriano Veloso, and Fabrício Benevenuto. Explainable machine learning for fake news detection. In *ACM Conference on Web Science, ACM WebSci*, pages 17–26, 2019.

[26] Ismael Santana, CN Ferreira, LBX Costa, MO Sóter, LML Carvalho, J de C. Albuquerque, MF Sales, AL Candido, FM Reis, Adriano Veloso, and KB Gomes. Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *Journal of Endocrinological Investigation*, 36(2):1–9, 2022.

[27] Roberta Viola, Guilherme Drummond, Adriano Veloso, and Mauricio Zuardi. A data-centric approach for predicting individual outcomes in a multi-party legislative system. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8. IEEE, 2022.

[28] Gianlucca Zuin, Rob Buechler, Tao Sun, Chad Zanocco, Daniella Castro, Adriano Veloso, and Ram Rajagopal. Revealing the impact of extreme events on electricity consumption in brazil: A data-driven counterfactual approach. In *IEEE Power & Energy Society General Meeting, PESGM*, pages 1–5. IEEE, 2022.

[29] Gianlucca L. Zuin, Felipe Marcelino, Lucas Borges, João Couto, Victor Jorge, Mychell Laurindo, Glaucio Barcelos, Márcio Cunha, Valdeci Alvarenga, Henrique Rodrigues, Paulo Balsamo, and Adriano Veloso. Predicting heating sliver in duplex stainless steels manufacturing through rashomon sets. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8. IEEE, 2021.

[30] Gianlucca L. Zuin, Adriano Veloso, João Cândido Portinari, and Nivio Ziviani. Automatic tag recommendation for painting artworks using diachronic descriptions. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8. IEEE, 2020.

## Conferências Nacionais – 2020 a 2023

[31] Geanderson E. dos Santos, Eduardo Figueiredo, Adriano Veloso, Markos Viggiato, and Nivio Ziviani. Predicting software defects with explainable machine learning. In *Brazilian Symposium on Software Quality, SBQS*, page 18, 2020.

[32] Geanderson E. dos Santos, Adriano Veloso, and Eduardo Figueiredo. Understanding thresholds of software features for defect prediction. In *Brazilian Symposium on Software Engineering SBES*, pages 305–310, 2022.

## Periódicos – 2020 a 2023

[33] Tiago Amador, Saulo Saturnino, Adriano Veloso, and Nivio Ziviani. Early identification of ICU patients at risk of complications: Regularization based on robustness and stability of explanations. *Artif. Intell. Medicine*, 128:102283, 2022.

[34] Daniella Castro Araújo, Adriano Veloso, Karina Braga Gomes Borges, and Maria das Graças Carvalho. Prognosing the risk of COVID-19 death through a machine learning-based routine blood panel: A retrospective study in brazil. *Int. J. Medical Informatics*, 165:104835, 2022.

[35] Daniella Castro Araújo, Adriano Veloso, Renato Santos de Oliveira Filho, Marie-Noelle Giraud, Leandro José Raniero, Lydia Masako Ferreira, and Renata Andrade Bitar. Finding reduced raman spectroscopy fingerprint of skin samples for melanoma diagnosis through machine learning. *Artif. Intell. Medicine*, 120:102161, 2021.

[36] Daniella Castro AraÃºjo, Adriano Veloso, Karina Braga Gomes, Leonardo Cruz de Souza, Nivio Ziviani, Paulo Caramelli, and Alzheimers Disease Neuroimaging Initiative. A novel panel of plasma proteins predicts progression in prodromal alzheimers disease. *Journal of Alzheimers Disease*, 88(2):549–561, 2022.

[37] Marcia CanÃ§ado, Luana Amaral, Evelin Amorin, Adriano Veloso, and Heliana Mello. Subjetividade em correções de redações. *Linguamatica*, 12(1):63–79, 2020.

[38] Anderson Bessa Da Costa, Larissa Moreira, Daniel Ciampi De Andrade, Adriano Veloso, and Nivio Ziviani. Predicting the evolution of pain relief: Ensemble learning by diversifying model explanations. *ACM Trans. Comput. Heal.*, 36(2):1–28, 2021.

[39] Geanderson Esteves dos Santos, Eduardo Figueiredo, Adriano Veloso, Markos Viggiato, and Nivio Ziviani. Understanding machine learning software defect predictions. *Autom. Softw. Eng.*, 27(3):369–392, 2020.

[40] Caio Jeronimo, Leandro Balby Marinho, Claudio Campelo, Adriano Veloso, and Allan Sales da Costa Melo. Characterization of fake news based on subjectivity lexicons. *J. Data Intell.*, 1(4):419–441, 2020.

[41] Larissa I Moreira, Anderson Bessa, N Ziviani, MJ Teixeira, J Rosi, M Nishio, Daniella Araujo, Ana Paula Macedo, GT Kubota, V Silva, Adriano Veloso, and Daniel Ciampi de Andrade. An artificial intelligence solution to detect potential non-response to chronic-pain treatment based on the first medical encounter. *EFIC2022 European Pain Federation*, 12:12–20, 2022.

[42] Julio C. S. Reis, André Correia, Fabricio Murai, Adriano Veloso, Fabrício Benevenuto, and Erik Cambria. Supervised learning for fake news detection. *IEEE Intell. Syst.*, 34(2):76–81, 2019.

[43] Alvaro Salgado, Raquel Melo-Minardi, M Giovanetti, Adriano Veloso, and F Morais-Rodrigues. Machine learning models exploring characteristic single-nucleotide signatures in yellow fever virus. *Plos One*, 17(12):e0278982, 2022.

[44] Ismael Santana, CN Ferreira, LBX Costa, MO Sóter, LML Carvalho, J de C. Albuquerque, MF Sales, AL Candido, FM Reis, Adriano Veloso, and KB Gomes. Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *Journal of Endocrinological Investigation*, 36(2):1–9, 2022.

[45] Marconi Santiago, Daniella Castro Araújo, Larissa R Stival, David Smadja, and Adriano Veloso. Ectasia risk model: A novel method without cut-off point based on artificial intelligence improves detection of higher-risk eyes. *Journal of Refractive Surgery*, 38(11):716–724, 2022.

[46] Ismael Santana Silva and Adriano Veloso. Automatic model evaluation using feature importance patterns on unlabeled data. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8. IEEE, 2022.

[47] A Veloso and N Ziviani. Explainable death toll motion modeling: COVID-19 data-driven narratives. *Plos One*, 17(4):e0264893, 2022.

[48] Gianlucca Zuin, Daniella Araujo, Vinicius Ribeiro, Maria Gabriella Seiler, Wesley Heleno Prieto, Maria Carolina PintÃ£o, Carolina dos Santos Lazari, Celso Francisco Hernandes Granato, and Adriano Veloso. Prediction of sars-cov-2-positivity from million-scale complete blood counts using machine learning. *Nature Comm. Medicine*, 2(1):72–90, 2022.

[49] Gianlucca Zuin, Rob Buechler, Tao Sun, Chad Zanocco, Francisco Galuppo, Adriano Veloso, and Ram Rajagopal. Extreme event counterfactual analysis of electricity consumption in brazil: Historical impacts and future outlook under climate change. *Energy*, 281(15):128101, 2023.

[50] Gianlucca L. Zuin, Luiz Chaimowicz, and Adriano Veloso. Deep learning techniques for explainable resource scales in collectible card games. *IEEE Trans. Games*, 14(1):46–55, 2022.

# HarmonyML: Aprendizado de modelos alinhados aos valores humanos e às restrições de negócio

## Sumário Executivo

**Objetivos:** O principal objetivo deste projeto é desenvolver métodos e algoritmos de Aprendizado de Máquina que façam uso de novos conceitos e avanços recentes relacionados à explicabilidade/interpretabilidade de forma a produzir modelos mais alinhados com valores humanos e/ou regras do negócio. Objetivos secundários incluem a exploração de diferentes cenários de aplicação, envolvendo tarefas complexas e de importância imediata, tais como a busca por exames de diagnóstico mais precisos (i.e., Alzheimer, COVID-19), a utilização mais eficaz de plantas energéticas (i.e., revezamento hidrotérmico), repropósito de drogas (i.e., remédios para COVID-19), entre outras. Serão realizadas parcerias tanto com a academia quanto com o setor privado a fim de se obter dados reais bem como proximidade com desafios práticos. Ressalta-se que para tais cenários de aplicação, é de suma importância a capacidade de se produzir modelos alinhados com o entendimento do fenômeno sendo modelado ou mesmo com as regras de negócio, ou seja, alinhados com explicações plausíveis, aceitáveis e coerentes com o conhecimento prévio.

**Motivação:** Uma grande ambição da comunidade de Aprendizado de Máquina é prover soluções a tarefas complexas que exigem predições com níveis de precisão próximos ou superiores à inteligência humana. Recentemente tal ambição vem sendo atingida em decorrência de avanços na área, evidenciando que o custo de se desenvolver modelos a partir de dados vem caindo muito rapidamente. A redução do custo, por sua vez, torna cada vez mais pervasiva a utilização de modelos preditivos/generativos, e por isso, é crescente a preocupação com os mecanismos por trás desses modelos. Questões como justiça, viés e até mesmo o alinhamento com a realidade conhecida, tornam-se agora cruciais para a adoção desses modelos. A motivação para nosso projeto, portanto, é a necessidade de se produzir modelos alinhados aos valores humanos e de negócio, de forma a garantir que os modelos produzidos sejam válidos, justos e aceitáveis.

**Descrição:** O projeto é organizado em três camadas. Na primeira camada iremos definir as aplicações de nosso interesse. As aplicações refletem uma diversidade de problemas relevantes e de alto impacto, e os dados correspondentes são obtidos através de parcerias estabelecidas com pesquisadores, tanto na academia quanto no setor privado. A segunda camada envolve a elaboração de novos algoritmos de aprendizado de máquina. Finalmente, na terceira camada iremos avaliar os algoritmos desenvolvidos.

**Resultados Esperados:** Ao fim deste projeto (36 meses) espera-se obter os seguintes resultados: (i) projetar, implementar e validar novos algoritmos que consigam produzir modelos (preditivos ou generativos) que sejam mais alinhados aos valores humanos e/ou aos negócios; (ii) avaliar a efetividade prática dos algoritmos desenvolvidos em tarefas relevantes e desafiadoras; (iii) formar 3 doutores, 8 mestres e 8 alunos de iniciação científica; (iv) publicar 6 artigos em periódicos, 10 artigos em conferências internacionais e 6 artigos em conferências nacionais; e (v) publicar e prover acesso ao estado-da-arte à

academia e empresas para que essas possam se beneficiar de oportunidades abertas pelo projeto proposto.

**Relação com o Projeto Anterior (Agosto 2020 – Julho 2023):** O projeto anterior foi focado no desenvolvimento de algoritmos de Aprendizado de Máquina que obedecem restrições em relação aos modelos preditivos que produzem. Tais restrições são aplicadas ao mecanismo de funcionamento dos modelos, e portanto, o projeto proposto pode ser considerado um avanço natural e esperado em relação ao projeto anterior. Além disso, vamos explorar outros desafios computacionais e cenários de aplicação, incluindo desafios relacionados ao Processamento de Linguagem Natural e à Visão Computacional.

**Recursos Solicitados:** Solicitamos a progressão para a bolsa de produtividade em pesquisa nível 1D, considerando a produção técnico-científica e atuação acadêmica do proponente. Em particular, o pedido de progressão é justificado já que o proponente atende ao perfil esperado na categoria 1D como descrito no Anexo I do Edital, a saber:

- tem mais de 8 anos de doutorado.
- apresenta produção com regularidade desde 2011.
- tem produções qualificadas em nível internacional, várias em periódicos, considerados de primeira linha.
- é vinculado ao Programa de Pós-Graduação do Departamento de Ciência da Computação da UFMG desde 2011, e já orientou dezenas de dissertações de mestrado e teses de doutorado com êxito.
- apresenta reconhecida liderança nacional nas áreas de Aprendizado de Máquina, Aprendizado Profundo, e Processamento de Linguagem Natural. Forma recursos humanos altamente qualificados, que assumem posições de destaque tanto no Mercado quanto na Acadêmia.
- demonstra capacidade de captar recursos financeiros e humanos para a realização de pesquisa de ponta em sua área de atuação.

# 1 Introdução

A modelagem explanatória, a modelagem preditiva e a modelagem generativa, são maneiras de construir abstrações úteis a partir de dados [70]. Embora haja uma sensação instintiva de que essas modelagens sejam tarefas diferentes, muitas vezes presume-se que as habilidades de predizer e gerar sejam conectadas e relacionadas à habilidade de explicar. Ainda assim, a maior parte da literatura recente não explora nenhum tipo de relação entre essas habilidades durante a construção de modelos a partir dos dados, o que por vezes nos leva a problemas devido à falta de alinhamento.

Alinhamento no contexto de modelos produzidos a partir dos dados refere-se ao conceito de garantir que esses modelos, particularmente aqueles altamente capazes ou avançados, se alinhem com nossos valores, objetivos e intenções. O objetivo do alinhamento é criar modelos de aprendizado de máquina que funcionem em harmonia com os valores humanos e com os objetivos do negócio, em vez de se comportar de maneiras que possam ser potencialmente prejudiciais ou desalinhadas com tais valores e objetivos. A tecnologia chegou em um momento de muita sofisticação e alta adoção no qual torna-se essencial garantir o alinhamento. À medida que os modelos de aprendizado de máquina tornam-se mais sofisticados e capazes, eles podem tomar decisões e ações com consequências de longo alcance. Se essas decisões e ações não estiverem alinhadas com os valores humanos e objetivos do negócio, elas podem levar a resultados negativos não intencionais.

Acreditamos que a busca pelo alinhamento deva ocorrer durante o desenvolvimento e evolução dos modelos. O desafio do alinhamento é particularmente relevante no desenvolvimento de sistemas avançados de Inteligência Artificial, como sistemas superinteligentes ou agentes altamente autônomos. Neste projeto, propomos trabalhar em técnicas e metodologias para garantir que esses sistemas permaneçam alinhados com os valores humanos e com as regras e objetivos do negócio ao longo de sua operação e evolução. Para tanto, os modelos precisam se adaptar a contextos em mudança, ser adaptáveis e capazes de lidar com novas situações ou cenários que não foram explicitamente antecipados durante a fase de design. Devem também ser capazes de evitar comportamentos indesejados ou prejudiciais, mesmo que não sejam explicitamente proibidos.

**Exemplo 1:** Um exemplo de falta de alinhamento que poderia causar sérios problemas para uma empresa está relacionado à interação de um chatbot de atendimento ao cliente com os clientes. Imagine que uma empresa desenvolva um chatbot de atendimento ao cliente com o objetivo de otimizar o suporte e melhorar a experiência do cliente. No entanto, devido a uma falta de alinhamento adequado, o chatbot é treinado com um conjunto de dados que não captura nuances culturais, emocionais e contextuais das interações com os clientes. Isso pode levar a problemas sérios:

- Desentendimento cultural: Se o chatbot não compreender adequadamente as diferenças culturais e linguísticas dos clientes, pode responder de maneira inadequada ou ofensiva, criando atrito e prejudicando as relações com os clientes.

- Respostas insensíveis: A falta de alinhamento emocional pode fazer com que o chatbot responda insensivelmente a problemas ou reclamações dos clientes, resultando em frustração e insatisfação.

- Incapacidade de resolver problemas complexos: Se o chatbot não estiver alinhado com os objetivos de oferecer soluções eficazes, ele pode não ser capaz de resolver problemas complexos dos clientes, resultando em repetidas interações e frustração.
- Perda de clientes e receita: A má experiência do cliente devido ao chatbot mal alinhado pode levar à perda de clientes e, consequentemente, à diminuição da receita.
- Desperdício de recursos: A empresa pode gastar recursos significativos no desenvolvimento e implantação do chatbot, apenas para enfrentar consequências negativas devido à falta de alinhamento.

**Exemplo 2:** Imagine que uma empresa do setor de metalurgia desenvolva um certo tipo de aço, com grande valor agregado. No entanto, a produção desse aço está em um momento de surto de um certo defeito. Um modelo preditivo é construído para identificar as possíveis fontes do defeito, e para além disso, o modelo também é capaz de propor ações para reduzir/impedir o defeito. No entanto, devido a uma falta de ainhamento adequado, o modelo preditivo acaba por propor ações impraticáveis, como por exemplo aumentar a concentração de um elemento químico de custo muito alto (i.e., molibidênio). Novamente, isso pode levar a problemas sérios:

- O produto produzido pode ficar fora de sua especificação: Isso poderia causar outros defeitos, conhecidos ou até mesmo desconhecidos.
- Aumento do custo: O elemento químico, em concentrações mais elevadas, irá acrescentar muito custo de produção ao produto.
- Desperdício de recursos: Se a ação é impraticável, os recursos despendidos para a criação e implementação do modelo preditivo foram parcialmente desnecessários.

Esses exemplos destacam a importância de garantir que os modelos estejam devidamente alinhados com os objetivos e valores da empresa para evitar consequências indesejadas. A seguir descrevemos o que acreditamos serem formas de buscar o alinhamento:

- Verificação de alinhamento: Pretendemos criar e usar métodos de verificação para avaliar se o modelo está agindo de acordo com os objetivos definidos. Isso pode envolver simulações e modelos explicativos [59, 68, 69], de modo que seja possível entender como os modelos tomam decisões. Isso ajudará a identificar desalinhamentos e corrigir os modelos.
- Envolvimento de especialistas: Pretendemos incluir especialistas, usuários e partes interessadas relevantes no processo de desenvolvimento para garantir que várias perspectivas sejam consideradas.
- Regularização e restrições: Pretendemos criar e implementar técnicas de regularização e restrições durante o treinamento dos modelos para garantir que o modelo não faça escolhas que violem valores ou objetivos críticos.

Nesse sentido, acreditamos existir uma relação mais virtuosa entre modelos preditivos, generativos e explanatórios, e que podemos tirar proveito de tal relação de forma a produzir modelos alinhados, que sejam mais confiáveis e justos. Em particular, neste projeto propomos controlar a busca por modelos preditivos[3] levando-se em conta restrições impostas quanto à importância esperada para cada característica ou hiperparâmetro empregado pelo modelo. Desta forma, assumimos a disponibilidade de uma visão (mesmo que parcial) em relação às explicações esperadas (ou corretas) [56], possibilitando que a busca por modelos seja otimizada de maneira a se obter modelos preditivos seguros, oferecendo justiça e/ou corretude e ao mesmo tempo preservando ao máximo a eficácia das predições. Com esta proposta, vislumbramos uma série de possíveis avanços, sumarizados a seguir:

- Do ponto de vista teórico, pretendemos oferecer uma visão alternativa a respeito do risco de sobreajuste dos modelos (overfitting), bem como sobre a capacidade de generalização dos modelos preditivos. Uma possível pergunta de pesquisa, neste caso, seria: há uma relação entre modelos preditivos sobreajustados e a impossibilidade de se respeitar as restrições de explicação impostas? Para responder tal pergunta pretendemos modelar fenômenos para os quais temos conhecimento prévio, mesmo que parcial, da cadeia causal.

- Do ponto de vista prático, uma série de novos algoritmos de busca por modelos serão propostos e avaliados. Tais algoritmos apresentarão diferenças no que diz respeito a como a busca por modelos será realizada. Mais especificamente, proporemos algoritmos que não permitem infringir restrições ou que permitem o relaxamento das restrições em troca de ganhos de eficácia. Acreditamos que cada algoritmo proposto oferecerá uma perspectiva diferente sobre a escolha entre justiça/corretude e eficácia.

Os algoritmos a serem propostos neste projeto serão avaliados em cenários de aplicação sofisticados e com grande potencial de inovação, muitas vezes explorando dados originais, o que exigirá a realização de parcerias com pesquisadores de outras disciplinas. Aplicações nas quais pretendemos avaliar nossos algoritmos incluem:

- Modelos de diagnóstico de COVID-19 através de hemograma. Para tanto firmamos parceria com o Grupo Fleury. Temos acesso a milhões de hemogramas, centenas de milhares de testes RT-PCR, e dezenas de milhares de testes de sorologia.

- Modelos para estimar o risco de óbito de pacientes intensivos. Para tanto firmamos parceria com o Hospital LifeCenter de Belo Horizonte. Temos acesso a aproximadamente 10 mil históricos de pacientes admitidos na UTI do hospital.

- Modelos para estimar a chance de melhora de pacientes com dor crônica. Para tanto firmamos parceria com a Faculdade de Medicina da Universidade de São Paulo. Temos acesso a quase mil pacientes acometidos de dor crônica que realizaram pelo menos uma consulta médica no Hospital das Clínicas de São Paulo.

[3] A busca acontece em termos de variações nos hiperparâmetros e da seleção de características que irão compor o modelo.

- Modelos para otimização de despacho energético através do revezamento hidrotérmico. Para tanto firmamos parceria com a empresa Power Systems Research. Temos acesso ao histórico de disponibilidade hídrica, bem como à configuração e capacidade técnica das usinas hidrelétricas e termelétricas da Colômbia.

- Modelos para estimar a efetividade de drogas contra a COVID-19. Para tanto firmamos parceria com o Centro Nacional de Pesquisa em Energia e Materiais. Temos acesso a centenas de propriedades sobre milhares de drogas.

- Modelos para identificar traços de dor na face de fetos em imagens de ultrassonografia. Para tanto firmamos parceria com a Maternidade Sepaco. Temos acesso a dezenas de vídeos mostrando os fetos antes e após serem submetidos à cirurgia.

É importante ressaltar que para todos esses cenários de aplicação, obtivemos acesso à pesquisa prévia de onde extraímos o conjunto de explicações esperadas e encaradas como corretas pelos especialistas, e que serão utilizadas como restrições a serem impostas aos mecanismos de predição e geração. Além disso, o proponente pretende que a condução da pesquisa seja amparada pela construção de pacotes de software que sirvam como bancada de teste para experimentos, obtendo resultados aferíveis através de publicações e formação de recursos humanos altamente qualificados nos níveis de iniciação científica, mestrado e doutorado. Posteriormente, alguns desses resultados poderão ser repassados para a indústria. A seguir apresentamos nossa visão abstrata sobre os objetivos listados acima através da exposição de resultados preliminares que motivam a pesquisa proposta. Também apresentaremos nosso trabalho pregresso, metodologia e objetivos específicos.

# 2 Modelos seguros *by design*

Este projeto de pesquisa versa sobre a elaboração, desenvolvimento e avaliação de novos algoritmos de Aprendizado de Máquina que façam uso de conceitos e avanços recentes relacionados à explicabilidade/interpretabilidade de forma a produzir modelos preditivos mais alinhados com o conhecimento prévio sobre o fenômeno sendo modelado. Atualmente, a busca por modelos em geral dá-se unicamente em termos da tentativa de escolher o modelo que forneça a menor expectativa de erro de predição (ou geração). No entanto, a comunidade vem observando de maneira crescente, que essa abordagem pode resultar em modelos eficazes, porém injustos ou incorretos [54, 55, 57, 58, 62, 64]. A literatura mais recente já contempla iniciativas demonstrando a possibilidade de se construir modelos a partir de explicações [60]. O desafio de interesse neste projeto é similar, e os algoritmos a serem desenvolvidos irão incorporar restrições de explicação definindo o conhecimento, mesmo que parcial, acerca do fenômeno sendo modelado, de forma a reduzir o espaço de busca por modelos focando naqueles cujos mecanismos de predição/geração sejam considerados seguros.

## 2.1 Restrições quanto ao mecanismo de predição/geração

A base de nosso projeto é a possibilidade de podermos associar um modelo explanatório a um modelo preditivo our generativo, de forma que o modelo explanatório revele o

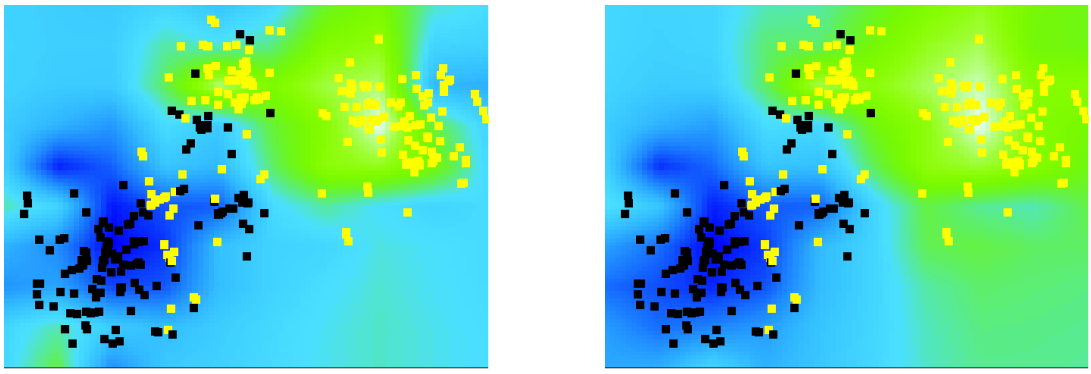

Figure 1: Separações produzidas por dois modelos preditivos diferentes. Na esquerda o modelo preditivo foi obtido sem uso de restrições. Na direita o modelo preditivo foi obtido empregando-se restrições nas características "cortisol" e "proteína C-reativa".

mecanismo de predição/geração que foi empregado. Tome como exemplo um modelo preditivo capaz de predizer se um paciente, sem sintomas, irá avançar para a fase sintomática da Doença de Alzheimer dentro dos próximos 4 anos. Suponha que o modelo preditivo seja construído a partir das concentrações observadas de algumas proteínas nos plasma sanguíneo do paciente. Nesse caso, a literatura relevante aponta que pacientes na fase sintomática da doença apresentam elevadas concentrações de cortisol e concentrações diminuídas de proteína C-reativa (CRP) [65, 66]. No que diz respeito ao nosso projeto, tal conhecimento prévio poderia ser incorporado ao processo de busca por modelos. Mais especificamente, o conhecimento prévio produziria restrições e mecanismos de predição/geração válidos precisariam estar em concordância com essas restrições. Sendo assim, dentre os muitos modelos que poderiam ser extraídos dos dados, só seriam considerados aqueles cujos mecanismos de predição/geração fossem válidos.

A Figura 1 mostra as separações produzidas por dois modelos preditivos diferentes. O espaço de entradas (i.e., pacientes) é dividido em duas regiões − a região verde corresponde a predições negativas (i.e., o paciente não irá avançar para a fase sintomática) e a região azul corresponde a predições positivas (i.e., o paciente irá avançar para a fase sintomática). Ainda na figura, pontos pretos correspondem a pacientes que avançaram para a fase sintomática da doença, enquanto pontos amarelos correspondem a pacientes que não avançaram. Os dois modelos apresentam erros de predição em termos de área abaixo da curva ROC [53] muito similares, porém produzem separações ligeiramente diferentes, como mostrado na figura. A Figura 2 mostra o modelo explanatório [59] que foi associado ao modelo preditivo selecionado.

Na figura, pontos vermelhos correspondem a altas concentrações da proteína correspondente, en quanto pontos azuis correspondem a baixas concentrações. O eixo $x$ mostra o efeito que a concentração observada tem no mecanismo de predição do modelo escolhido. Efeito positivo contribui para aumentar a probabilidade de avanço, e da mesma forma o efeito negativo contribui para diminuir a probabilidade de avanço. Dessa forma, podemos constatar que o mecanismo de predição está em concordância com as restrições empregadas, uma vez que concentrações elevadas de cortisol contribuem para o aumento da probabilidade predição positiva, e a tendência contrária ocorre com a proteína C-reativa. Certamen-

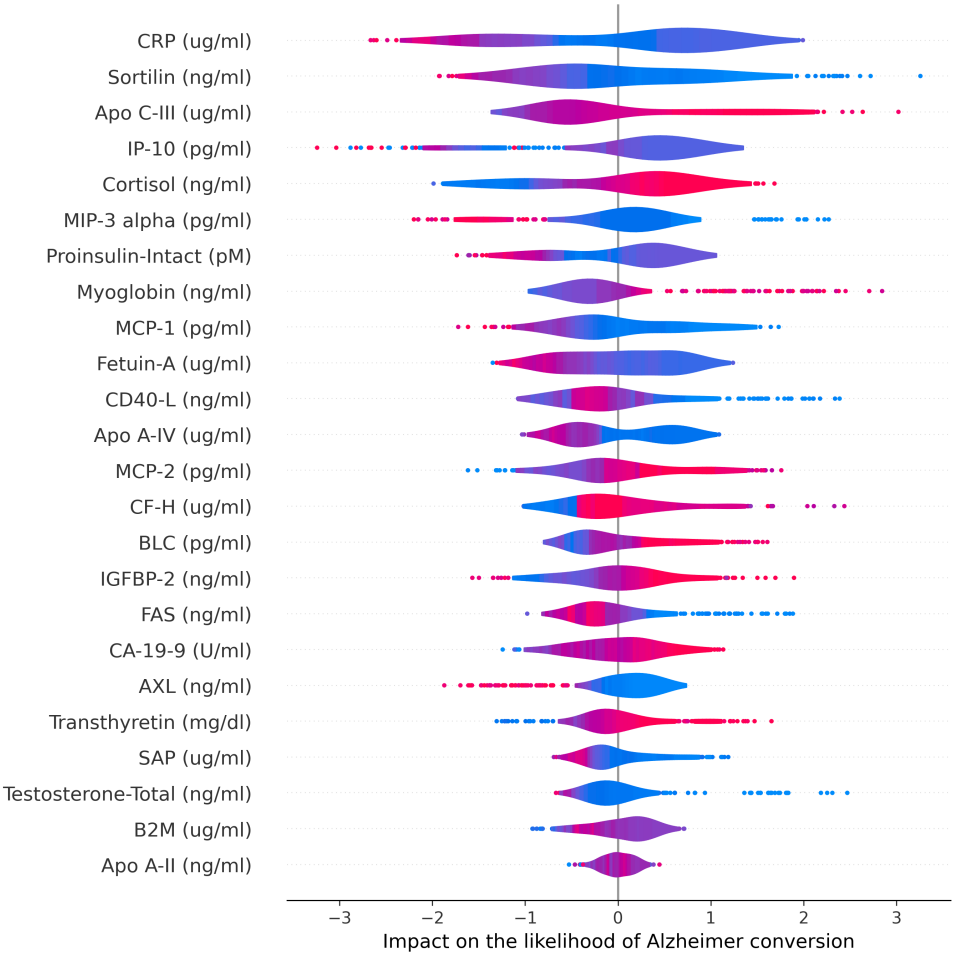


Figure 2: Mecanismo de predição empregado pelo modelo obtido empregando-se restrições nas características "cortisol" e "proteína C-reativa".

te, há várias maneiras de criarmos restrições com base no conhecimento prévio sobre o fenômeno sendo modelado. Embora o exemplo acima não apresente detalhes de como as restrições possam ser criadas, esse tópico é um dos objetivos específicos deste projeto.

## 2.2 Busca no espaço de modelos

No geral, os algoritmos de Aprendizado de Máquina requerem a sintonização de um conjunto de hiperparâmetros, os quais essencialmente ditam ao algoritmo como combinar o conjunto de características disponíveis [61]. Dessa forma, um modelo preditivo/generativo é o resultado da aplicação do algoritmo dada uma sintonização e um conjunto de características a serem combinadas.

Na prática, o espaço de possíveis modelos pode ser explorado variando-se o valor dos hiperparâmetros, bem como selecionando as características desejadas para compor o modelo [51]. A busca no espaço de possíveis modelos geralmente ocorre exclusivamente através da minimização do erro de predição/geração esperado, resultando assim em mode-

los com aparente baixo erro esperado. Ao empregar restrições durante o processo de busca por modelos, estamos diminuindo o número de modelos a serem considerados. Basicamente, o objetivo continua sendo encontrar o modelo preditivo/generativo com o menor erro esperado, porém dentro do espaço de modelos com mecanismos de predição/geração considerados válidos dado o conjunto de restrições. Mais especificamente, a toda vez que um modelo candidato for obtido, será construído também um modelo explanatório de forma a evidenciar o mecanismo de predição/geração do modelo. Dessa forma, as restrições podem ser comparadas ao mecanismo de predição/geração, validando ou invalidando o modelo candidato.

Dependendo da complexidade do fenômeno sendo modelado, é comum que sejam encontrados diversos modelos com desempenho similar, mas que foram obtidos com sintonizações e características diferentes. Por vezes, a diferença de desempenho é negligenciável. Dada a multiplicidade de modelos com desempenho similares, acreditamos que, caso as restrições reflitam o mecanismo correto, o desempenho dos modelos preditivos/generativos encontrados não deve ser inferior ao desempenho do modelo encontrado adotando uma busca sem restrições. Mais ainda, um fator importante a ser investigado neste projeto diz respeito ao impacto das restrições na generalização dos modelos. Pretendemos assim, comparar modelos preditivos/generativos obtidos com e sem o uso de restrições e compará-los em termos de suas capacidades de generalização.

Tome como exemplo a busca por modelos capazes de diagnosticar pacientes em relação à COVID-19. Nesse caso, os modelos preditivos são produzidos a partir de aproximadamente 500 mil hemogramas. A Figura 3 mostra o desempenho esperado através de validação-cruzada, tanto em termos de área sob a curva ROC quanto em relação à precisão média. A figura também mostra o mecanismo de predição do modelo, através do qual pode-se constatar a grande importância da característica "idade". O conhecimento prévio, no entanto, aponta para independência do diagnóstico em relação à idade, e por isso, adicionamos a restrição apropriada. A busca com a restrição adicionada resultou no modelo mostrado na Figura 4, o qual demonstra um desempenho esperado ligeiramente inferior ao modelo encontrado com a busca sem restrições. Como observado no mecanismo de predição do novo modelo, a impossibilidade de respeitar a restrição fez com que a característica "idade" não fosse utilizada no novo modelo.

Posteriormente, avaliamos ambos os modelos preditivos em 200 mil hemogramas adicionais. Embora os desempenhos esperados para ambos os modelos sejam similares, foi interessante observar que os desempenhos obtidos ao aplicar os modelos nos 200 mil hemogramas adicionais foram bastante discrepantes. O modelo obtido sem restrições atingiu uma área sob a curva de significativamente menor do que a prevista, enquanto o desempenho do modelo obtido ao empregar a restrição na característica "idade" permaneceu bem mais próximo à estimativa por validação-cruzada. Pretendemos buscar novos exemplos como esse, e ao estudá-los também esperamos formatar uma nova teoria a respeito da capacidade de generalização desses modelos.

## 2.3 Restrições quanto ao efeito das características

O efeito de uma característica é definido como a relação que existe entre o valor assumido pela característica e o impacto dela na predição (ou geração). Sendo assim, dizemos que uma característica tem efeito negativo na predição se ela contribui para dimunir o valor

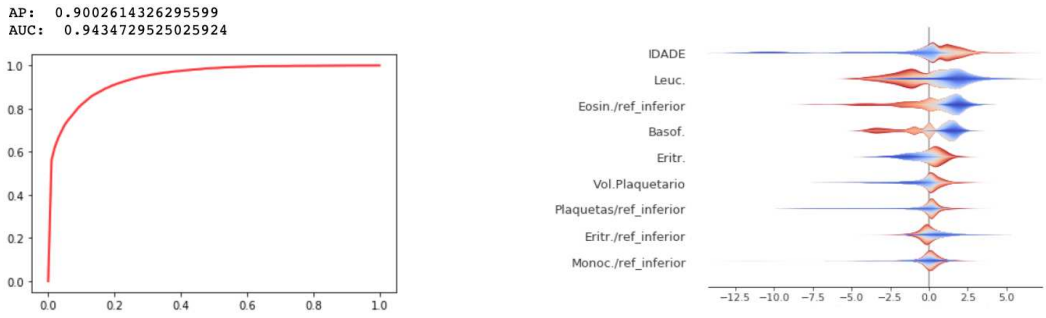


Figure 3: À esquerda, o desempenho esperado para o modelo. À direita, o mecanismo de predição do modelo.

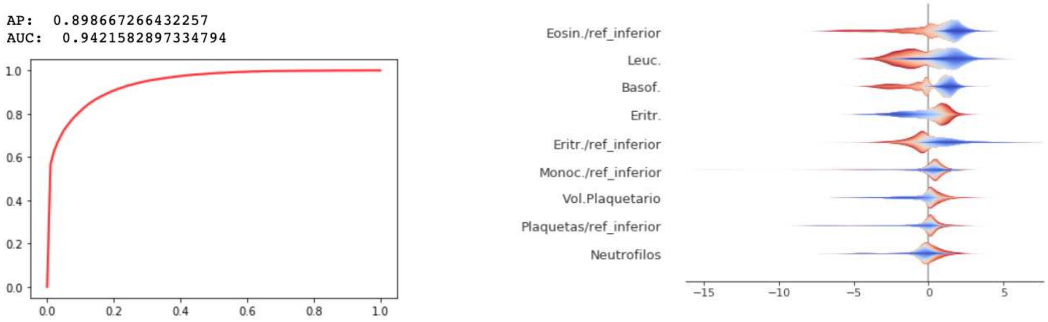


Figure 4: À esquerda, o desempenho esperado para o modelo. À direita, o mecanismo de predição do modelo.

da predição. Da mesma forma, dizemos que uma característica tem efeito positivo se ela contribui para aumentar o valor da predição. Dessa forma, as restrições serão dadas em termos dos efeitos esperados para cada característica, ou seja, para cada característica define-se o impacto negativo ou o impacto positivo esperados. Por fim, um relaxamento pode ser empregado ao considerarmos uma tolerância durante o processo de busca por modelos preditivos.

**Trabalho Pregresso.** Temos nos dedicado à assimilar algoritmos existentes e propor novos algoritmos que produzem modelos explanatórios. Em [67] avaliamos modelos explanatórios para expor mecanismos de propagação de notícias falsas pelo Facebook. Já em [52] propusemos um modelo explanatório para séries temporais, o qual foi utilizado para explicar predições acerca da evolução do quadro cínico de pacientes intensivos. No contexto de Aprendizado Profundo, propomos um novo algoritmo que produz explicações sobre a evolução da Doença de Alzheimer a partir de imagens de ressonância magnética [63]. Ainda no contexto de Aprendizado Profundo, propusemos uma medida [71] para a avaliação e comparação da qualidade das explicações fornecidas por expliacadores típicos. Sendo

assim, dada nossa experiência em construir modelos preditivos, generativos e explanatórios, é natural propor a exploração de possíveis relações entre esses tipos de modelagem de forma a obter modelos mais seguros, confiáveis e eficazes.

**Metodologia.** Para o desenvolvimento de nossos algoritmos utilizaremos os pacotes shap, scikit-learn, gpt4all, e Pytorch (dentre outros). O pacote shap implementa uma metodologia ótima de atribuição de importância às características. O pacote scikit-learn é o pacote padrão para construção de modelos de aprendizado de máquina. O pacote gpt4all fornece acesso a possibilidade de geração de textos e outras sequências. O Pytorch oferece suporte para reuso de estruturas e otimização em GPUs. A avaliação de nosso algoritmos sempre se dará com base na realização de experimentos controlados, com resultados comparados ao estado-da-arte, de acordo com medidas de eficácia apropriadas.

**Objetivos Específicos.** Os principais objetivos são:

- Aprimorar, estender e propor novos algoritmos de Aprendizado de Máquina que produzam modelos que apresentem alinhamento entre seu mecanismo de funcionamento e os objetivos do negócio. Alunos de mestrado e doutorado estarão envolvidos neste objetivo.
- Elaborar novas soluções baseadas em nossos algoritmos para os cenários de aplicação mencionados anteriormete. Alunos de mestrado e doutorado estarão envolvidos neste objetivo.
- Avaliar os algoritmos propostos, discutir e divulgar os resultados alcançados. Contamos com a participação de alunos de iniciação científica neste objetivo.

# 3 Resultados

Ao fim deste projeto (36 meses) espera-se obter os seguintes resultados:

- Projetar, implementar e validar novos algoritmos para o apredizado de modelos com alinhamento.
- Avaliar a efetividade prática dos algoritmos desenvolvidos em aplicações relevantes e desafiadoras.
- Formar 3 doutores, 8 mestres e 8 alunos de iniciação científica.
- Publicar 6 artigos em periódicos, 10 artigos em conferências internacionais e 6 artigos em conferências nacionais.
- Publicar e transferir a tecnologia produzida durante o projeto para fins de pesquisa e desenvolvimento tecnológico.

# 4 Recursos

Nesta seção discutimos a demanda e disponibilidade de recursos necessários para a execução do projeto proposto.

## 4.1 Bolsa de Produtividade

Espera-se a progressão para a categoria 1D, a qual é um pilar fundamental para a execução do projeto.

## 4.2 Recursos de Pessoal

Os demais recursos de pessoal para a realização do projeto estão disponíveis. Os alunos que trabalham nas linhas de pesquisa já estão cursando doutorado, mestrado ou atuando como bolsistas de iniciação científica. Acreditamos que eventuais substituições não afetarão significativamente o trabalho.

## 4.3 Recursos de Equipamento

Em termos de equipamentos, acreditamos que estejamos em condições de suprir as demandas de desenvolvimento e avaliação inerentes ao projeto. A infra-estrutura do laboratório LIA (Laboratório de Inteligência Artificial, sediado no DCC-UFMG e coordenado pelo proponente) foi recentemente renovada e estendida com recursos de projetos.

# References


[51] Ethem Alpaydin. *Introduction to Machine Learning*. Mit Press, 2014.

[52] Tiago Alves, Alberto H. F. Laender, Adriano Veloso, and Nivio Ziviani. Dynamic prediction of ICU mortality risk using domain adaptation. In *IEEE International Conference on Big Data, Big Data*, pages 1328–1336, 2018.

[53] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval, the concepts and technology behind search*. Pearson Education Ltd., 2011.

[54] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K. Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Conference on Fairness, Accountability, and Transparency, FAT**, pages 160–169, 2019.

[55] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Conference on Fairness, Accountability, and Transparency, FAT**, pages 79–88, 2019.

[56] Daniel Deutch and Nave Frost. Constraints-based explanations of classifications. In *35th IEEE International Conference on Data Engineering, ICDE*, pages 530–541, 2019.

[57] Severin Engelmann, Mo Chen, Felix Fischer, Ching-yu Kao, and Jens Grossklags. Clear sanctions, vague rewards: How china's social credit system currently defines "good" and "bad" behavior. In *Conference on Fairness, Accountability, and Transparency, FAT**, pages 69–78, 2019.

[58] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Conference on Fairness, Accountability, and Transparency, FAT**, pages 269–278, 2019.

[59] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NIPS*, pages 4765–4774, 2017.

[60] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT**, pages 1–9, 2019.

[61] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[62] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Conference on Fairness, Accountability, and Transparency, FAT**, pages 359–368, 2019.

[63] Eduardo Nigri, Nivio Ziviani, Fabio A. M. Cappabianco, Augusto Antunes, and Adriano Veloso. Explainable deep cnns for mri-based diagnosis of alzheimers disease. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2020.

[64] Ziad Obermeyer and Sendhil Mullainathan. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Conference on Fairness, Accountability, and Transparency, FAT**, page 89, 2019.

[65] Sid E. O'Bryant, Stephen C. Waring, Valerie Hobson, James R. Hall, Carol B. Moore, Teodoro Bottiglieri, Paul Massman, and Ramon Diaz-Arrastia. Decreased c-reactive protein levels in alzheimer disease. *J Geriatr Psychiatry Neurol*, 23(1):49–53, 2011.

[66] Sami Ouanes and Julius Popp. High cortisol and the risk of dementia and alzheimers disease: A review of the literature. *Front Aging Neurosci*, 11(42):1–11, 2019.

[67] Julio C. S. Reis, André Correia, Fabricio Murai, Adriano Veloso, and Fabrício Benevenuto. Explainable machine learning for fake news detection. In *ACM Conference on Web Science, ACM WebSci*, pages 17–26, 2019.

[68] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[69] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *The 32nd AAAI Conference on Artificial*

*Intelligence, the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 1527–1535, 2018.

[70] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[71] Dan Valle, Tiago Pimentel, and Adriano Veloso. Assessing the reliability of visual explanations of deep models with adversarial perturbations. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2020.