Adriano Alonso Veloso

e-mail: adrianov@dcc.ufmg.br

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

SafeML: Aprendizado de modelos preditivos seguros com restrições quanto ao mecanismo de predição

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

Projeto Anterior (Agosto 2017 – Julho 2020)

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 de Aprendizado Profundo baseados em arquiteturas de redes recorrentes, convolucionais, adversariais, sequência-sequência, treinadas por reforço ou através de supervisão e adaptação de domínio; (ii) avaliar a efetividade prática dos algoritmos desenvolvidos em tarefas relevantes e desafiadoras relacionadas ao Processamento de Linguagem Natural e Visão Computacional; (iii) formar 2 doutores, 8 mestres e 8 alunos de iniciação científica; (iv) publicar 6 artigos em periódicos e 10 artigos em conferências internacionais; e (v) publicar e transferir toda a tecnologia produzida durante o projeto para fins de pesquisa e desenvolvimento tecnológico.

Resultados Alcançados: (i), (ii) e (v) métodos e algoritmos de aprendizado profundo foram projetados, implementados e validados. 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; (iii) duas alunas (Evelin Amorin e Mariane Souza) tiveram suas teses aprovadas [3, 12], e um aluno (Anderson Bessa) passou pela qualificação de tese [4] e será doutor até 2021. Formamos 14 mestres (Lucas Miranda [11], Alison Marczewski [10], André Lloyd [8], Vinícius Melo [16], Alberto Albuquerque [1], Gianlucca Zuin [17], Dan Nascimento [15], Túlio Loures [9], Guillermo Contreras [6], Alex Barros [7], Tiago Alves [2], Tiago Pimentel [14], Thiago Oliveira [13], Dehua Chen [5]), e 10 alunos de iniciação científica (Igor Marques, Rafael Castro, João Marcos Couto, Breno Tanure, Luiz Melo, Daniel Augusto, Roberta Viola, Silvia Guerra, Lucas Lage, Lucas Aquino); (iv) publicamos 6 artigos em periódicos [43–48], 25 artigos em conferências internacionais [18, 20–29, 31–35, 37, 39–42, 50, 63, 67, 74] (outros 11 artigos estão submetidos e em processo de avaliação por pares com desfecho nos próximos meses).

Atuação no Período – 2017 a 2020

De acordo com a Chamada CNPq No 09/2020 – 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? Qual a importância dessas publicações na área de pesquisa da pós-graduação em questão? Quantas orientações de mestrado e doutorado estão em andamento? Que tipos de cursos de pós-graduação relacionados à sua pesquisa o propo-*

nente tem lecionado? Com que regularidade? Em que tipo de programa ou circunstância (e.g. cursos convidados em outras instituições, tutoriais em eventos relevantes, etc.)?"

Assim sendo, apresentamos a seguir, de forma mais detalhada, indicadores objetivos a respeito da atuação acadêmica do proponente no período relevante para o julgamento (2017 a 2020).

Orientações Concluídas: Foram defendidas 2 teses de doutorado e 14 dissertações de mestrado, detalhadas a seguir.

1. Mariane Moreira – Fashion retrieval in a semantic space: Balancing identity and fashionability, 08/02/2018. Banca examinadora: Adriano Veloso (UFMG, orientador), Leandro Balby (UFCG), Marco Cristo (UFAM), Rodrygo Santos (UFMG), Wagner Meira Jr. (UFMG). Artigos publicados: [61].

2. Evelin Amorim – Explainable models for automated essay scoring in the presence of biased scores, 16/12/2019. Banca examinadora: Adriano Veloso (UFMG, orientador), Márcia Cançado (UFMG, coorientadora), Fabricio Benevenuto (UFMG), Pedro Olmo (UFMG), Cilene Nevins (PUC-RJ), Helena Caseli (UFSCAR). Artigos publicados: [20, 21], citePmarcia.

3. Vinícius Melo – JSPY: um modelo objetivo para compreensão de linguagem natural, 17/03/2017. Banca examinadora: Adriano Veloso (UFMG, orientador), Adriano Pereira (UFMG), Fernando Magno (UFMG). Artigo submetido e em avaliação por pares.

4. André Lloyd – Extending Markov models through gradient descent optimization, 30/03/2017. Banca examinadora: Adriano Veloso (UFMG, orientador), Leandro Balby (UFCG), Renato Assunção (UFMG), Renato Ferreira (UFMG). Artigo submetido e em avaliação por pares.

5. Alberto Albuquerque – Recodificação de atributos para learning to rank usando autoencoders, 23/06/2017. Banca examinadora: Renato Ferreira (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Edleno de Moura (UFAM), Leandro Balby (UFCG), Nivio Ziviani (UFMG). Artigos publicados: [18].

6. Alison Marczewski – Learning transferable features from multiple source domains for speech emotion recognition, 17/07/2017. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), Hani Camille (UFMG). Artigos publicados: [28].

7. Gianlucca Zuin – Sistemas de Pergunta-Resposta de Grão-Fino baseados em Adaptação de Domínio e Informação Contextual, 14/11/2017. Banca examinadora: Adriano Veloso (UFMG, orientador), Luiz Chaimowics (UFMG, coorientador), Agma Traina (USP), Nivio Ziviani (UFCG), Renato Assunção (UFMG). Artigos publicados: [40]

8. Tiago Pimentel – Fast Node Embeddings: Learning Ego-Centric Representations, 26/02/2018. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), Rodrigo Barros (PUC-RS), Fabricio Murai (UFMG). Artigos publicados: [31–35].

9. Tiago Alves – Dynamic prediction of ICU mortality risk using domain adaptation, 28/02/2018. Banca examinadora: Adriano Veloso (UFMG, orientador), Alberto Laender (UFMG, coorientador), Nivio Ziviani (UFMG), Caetano Traina Jr. (USP), José Mauro Vieira Jr. (Hospital Sírio-Libanês). Artigos publicados: [50].

10. Dan Valle – Assessing the Reliability of Visual Explanations of Deep Models through Adversarial Perturbation, 27/03/2019. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG), William Schwartz (UFMG), Eduardo Valle Jr. (Unicamp). Artigos publicados: [39, 74].

11. Dehua Chen – Modeling Pharmacological Effects with Multi-Relation Unsupervised Graph Embedding, 30/03/2020. Banca examinadora: Adriano Veloso (UFMG, orientador), Nivio Ziviani (UFMG, coorientador), Deborah Schechtman (USP), Raquel Minardi (UFMG). Artigos publicados: [22, 23].

12. Túlio Loures – Representação de Entidades Baseada em Discussões, 10/10/2017. Banca examinadora: Pedro Olmo (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Wagner Meira Jr. (UFMG), Flavio Figueiredo (UFMG). Artigos publicados: [27], [46].

13. Alex Barros – A Flexible Compositional Approach to Word Sense Disambiguation, 27/07/2018. Banca examinadora: Nivio Ziviani (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Flavio Figueiredo (UFMG), Renato Ferreira (UFMG), Waldmir Cardoso Brandão (PUC-MG). Artigo submetido e em avaliação por pares.

14. Guillermo Contreras – IOT framework to improve productivity in open space offices, 29/04/2019. Banca examinadora: Daniel Macedo (UFMG, orientador), Adriano Veloso (UFMG, coorientador), José Marcos Nogueira (UFMG), Gilberto Medeiros (UFMG), Wladmir Cardoso Brandão (PUC-MG). Artigo submetido e em avaliação por pares.

15. Thiago Oliveira – Automatic Pain Assessment in Fetuses through Transfer Learning, 31/03/2020. Banca examinadora: Nivio Ziviani (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Flavio Figueiredo (UFMG), Daniel Ciampi (USP). Artigo submetido e em avaliação por pares.

16. Lucas Miranda – Manutenção Automática de um Sistema de Auditoria Inteligente em uma Seguradora de Saúde, 01/04/2020. Banca examinadora: Nivio Ziviani (UFMG, orientador), Adriano Veloso (UFMG, coorientador), Wagner Meira Jr. (UFMG), Adriano Pereira (UFMG). Artigo submetido e em avaliação por pares.

Orientações em Andamento: Há 9 orientações de doutorado e 6 orientações de mestrado em curso, detalhadas a seguir.

1. Ricardo Alves, doutorado (orientador), defesa prevista para março 2025.

2. Isamel Santana, doutorado (orientador), defesa prevista para agosto de 2024.

3. Daniella Castro, doutorado (orientador), defesa prevista para março de 2024.

4. Geanderson Esteves, doutorado (coorientador), defesa prevista para março de 2023.
5. Gianlucca Zuin, doutorado (orientador), defesa prevista para março de 2023. Artigos publicados: [41, 42].
6. Alison Marczewski, doutorado (orientador), defesa prevista para novembro de 2022.
7. Amir Jalilifard, doutorado (orientador), defesa prevista para novembro de 2021. Artigos publicados: [23].
8. Anderson Bessa, doutorado (orientador), defesa prevista para novembro de 2021.
9. Tiago Amador, doutorado (orientador), defesa prevista para novembro de 2021. Artigos publicados: [18].
10. Lucas Parreiras, mestrado (orientador), defesa prevista para setembro de 2022.
11. Silvia Guerra, mestrado (orientador), defesa prevista para setembro de 2022.
12. Guilherme Mendes, mestrado (orientador), defesa prevista para março de 2022.
13. Felipe Glicério, mestrado (orientador), defesa prevista para março de 2021.
14. Victor Rodrigues Jorge, mestrado (orientador), defesa prevista para setembro de 2020.
15. Eduardo Nigri, mestrado (orientador), defesa prevista para setembro de 2020. Artigos publicados: [63].

Supervisões de Pós-Doutorado: Foi concluída uma supervisão de pós-doutorado e outra supervisão continua em curso:

1. Solange Mata Machado, Impacto do mindset na geração de projetos de inteligência artificial. Início: 2017/02. Encerramento: 2019/02
2. Raquel Fabreti de Oliveira, Aplicação de metodologias da inteligência artificial para predizer fatores de risco e sobrevida no transplante renal. Início: 2020/01

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. Plataforma Big Data para Computação Cognitiva – Projeto de Pesquisa e Desenvolvimento em Parceria com a empresa Kunumi. Valor: R$180.000,00.
3. Antecipação de Qualidade na Produção de Aço Inoxidável – Projeto de Pesquisa e Desenvolvimento em Parceria com a empresa Aperam South America. Valor: R$392.299,96

4. Solução Inteligente para Gestão em Terapia Intensiva − Projeto de Pesquisa e Desenvolvimento em Parceria com a empresa Kunumi. Valor: R$503.436,46

5. Aprendizado Novos Algoritmos e Arquiteturas para Aprendizado Profundo, 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. MASWEB, Modelos, Algoritmos e Sistemas para a Web − CNPq/Fapemig/Pronex. Valor: R$330.000,00.

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

3. Modelagem do comportamento do parlamentar por meio de inteligência artificial. Projeto de Pesquisa e Desenvolvimento com a empresa Dado Capital. Valor: R$783.944,13

4. INWeb, Instituto Nacional de Ciência e Tecnologia para a Web − MCT/CNPq. Valor: R$2.300.000,00.

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$10.253.836,76, dos quais R$2.765.336,42 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).

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 Mineração de Dados:

- 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,

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

- 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,

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. Por fim, o proponente publica aproximadamente um artigo em periódico a cada três artigos em evento.

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 – 2017 a 2020

[1] Alberto Albuquerque. Recodificação de atributos para learning to rank usando autoencoders (Mestrado), 2017.

[2] Tiago Alves. Dynamic prediction of icu mortality risk using domain adaptation (Mestrado), 2018.

[3] Evelin Amorim. Explainable modelos for automated essay scoring in the presence of biased scores (Doutorado), 2019.

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

Indicador	**5 anos**	**10 anos**
H-Index (*Google Scholar*)	22	28
Citações (*Google Scholar*)	1550	2620
Livros	0	1
Periódicos	11	22
Artigos em conferências	30	59
Orientações concluídas	36	62
Doutorado	2	3
Mestrado	16	30
Iniciação Científica	18	29
Orientações em andamento	25	–
Doutorado	9	–
Mestrado	6	–
Iniciação Científica	10	–
Projetos de pesquisa na própria instituição	8	16
Coordenação	5	8
Participação	3	8
Projetos de pesquisa multi-institucional	3	6
Participação	3	6
Projetos de pesquisa de cooperação internacional	1	3
Participação	1	3
Projetos de pesquisa com org. públicas ou privadas	5	11
Coordenação	2	4
Participação	3	7
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

Tabela 1: Sumário com indicadores de desempenho em pesquisa.

	Eventos		**Periódicos**	
	5 anos	**10 anos**	**5 anos**	**10 anos**
A1	16	29	5	10
A2	5	8	1	2
B1	4	9	3	5

Tabela 2: Publicações em eventos e periódicos qualificados.

Tabela 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)
2020.1	Aprendizado de Máquina (PG)	34 (grad) + 50 (pós)
2010.1−2020.1		363 (grad) + 484 (pós)

[4] Anderson Bessa. Explainability, predictability and counterfactuals: An alternative to all-in-one approach (Doutorado), 2019.

[5] Dehua Chen. Modeling pharmacological effects with multi-relation unsupervised graph embedding (Mestrado), 2020.

[6] Guillermo Contreras. A smart iot office employing machine learning for personalised user comfort (Mestrado), 2019.

[7] Alex de Paula Barros. A flexible compositional approach to word sense disambiguation (Mestrado), 2018.

[8] André Lloyd Dwight Perlee Harder. Extending markov models through gradient descent optimization (Mestrado), 2017.

[9] Túlio Loures. Representação de entidades baseada em discussões (Mestrado), 2017.

[10] Alison Marczewski. Learning transferable features from multiple source domains for speech emotion recognition (Mestrado), 2017.

[11] Lucas Miranda. Manutenção automática de um sistema de auditoria inteligente em uma seguradora de saúde (Mestrado), 2019.

[12] Mariane Moreira. Fashion retrieval in a semantic space: Balancing identity and fashionability (Doutorado), 2018.

[13] Thiago Oliveira. Automatic pain assessment in fetuses through transfer learning (Mestrado), 2020.

[14] Tiago Pimentel. Fast node embeddings: Learning ego-centric representations (Mestrado), 2018.

[15] Dan Valle. Assessing the reliability of visual explanations of deep models through adversarial perturbation (Mestrado), 2019.

[16] Vinícius Veloso. JSPY: um modelo objetivo para compreensão de linguagem natural (Mestrado), 2017.

[17] Gianlucca Zuin. Sistemas de pergunta-resposta de grão-fino baseados em adaptação de domínio e informação contextual (Mestrado), 2017.

Conferências Internacionais – 2017 a 2020

[18] Alberto Albuquerque, Tiago Amador, Renato Ferreira, Adriano Veloso, and Nivio Ziviani. Learning to rank with deep autoencoder features. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2018.

[19] 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.

[20] Evelin Amorim, Márcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 229–237, 2018.

[21] Evelin Amorim and Adriano Veloso. A multi-aspect analysis of automatic essay scoring for brazilian portuguese. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 94–102, 2017.

[22] Dehua Chen, Tiago Pimentel Martins da Silva, Adriano Alonso Veloso, Nivio Ziviani, and Wladmir Cardoso Brandao. Denoising node embeddings. In *Official LXAI Research Workshop (LXAI@Neurips)*, 2018.

[23] 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 1–8, 2020.

[24] André Costa, Roberto Nalon, Wagner Meira Jr., and Adriano Veloso. Ego-centric analysis of supportive networks. In *ACM Conference on Web Science, ACM WebSci*, pages 281–285, 2018.

[25] 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.

[26] Caio Libânio Melo Jerônimo, Leandro Balby Marinho, Cláudio E. C. Campelo, Adriano Veloso, and Allan Sales da Costa Melo. Fake news classification based on subjective language. In *International Conference on Information Integration and Web-based Applications & Services, iiWAS*, pages 15–24, 2019.

[27] Túlio C. Loures, Pedro O. S. Vaz de Melo, and Adriano Alonso Veloso. Generating entity representation from online discussions: Challenges and an evaluation framework. In *Brazillian Symposium on Multimedia and the Web, Webmedia*, pages 197–204, 2017.

[28] Alison Marczewski, Adriano Veloso, and Nivio Ziviani. Learning transferable features for speech emotion recognition. In *ACM Multimedia Conference, ACM MM*, pages 529–536, 2017.

[29] Antoni Mauricio, Fábio A. M. Cappabianco, Adriano Veloso, and Guillermo Cámara. A sequential approach for pain recognition based on facial representations. In *International Conference on Computer Vision Systems, ICVS*, pages 295–304, 2019.

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

[31] Tiago Pimentel, Rafael Castro, Adriano Veloso, and Nivio Ziviani. Efficient estimation of node representations in large graphs using linear contexts. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2019.

[32] 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.

[33] Tiago Pimentel, Adriano Veloso, and Nivio Ziviani. Unsupervised and scalable algorithm for learning node representations. In *International Conference on Learning Representations, ICLR*, 2017.

[34] Tiago Pimentel, Adriano Veloso, and Nivio Ziviani. Fast node embeddings: Learning ego-centric representations. In *International Conference on Learning Representations, ICLR*, 2018.

[35] Tiago Pimentel, Juliano Viana, Adriano Veloso, and Nivio Ziviani. Fast and effective neural networks for translating natural language into denotations. In *International Symposium on String Processing and Information Retrieval, SPIRE*, pages 334–347, 2018.

[36] 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.

[37] Allan Sales, Leandro Balby, and Adriano Veloso. Media bias characterization in brazilian presidential elections. In *ACM Conference on Hypertext and Social Media, ACM HT*, pages 231–240, 2019.

[38] 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.

[39] Dan Valle, Nivio Ziviani, and Adriano Veloso. Effective fashion retrieval based on semantic compositional networks. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2018.

[40] Gianlucca L. Zuin, Luiz Chaimowicz, and Adriano Veloso. Learning transferable features for open-domain question answering. In *International Joint Conference on Neural Networks, IJCNN*, pages 1–8, 2018.

[41] Gianlucca L. Zuin and Adriano Veloso. Learning a resource scale for collectible card games. In *IEEE Conference on Games, CoG*, pages 1–8, 2019.

[42] 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, 2020.

Periódicos – 2017 a 2020

[43] Aline Bessa, Rodrygo L. T. Santos, Adriano Veloso, and Nivio Ziviani. Exploiting item co-utility to improve collaborative filtering recommendations. *J. Assoc. Inf. Sci. Technol.*, 68(10):2380–2393, 2017.

[44] Márcia Cançado, Luana Amaral, Evelin Amorin, Adriano Veloso, and Heliana Mello. Subjetividade em correções de redações detecção automática através de léxico de operadores de viés linguístico. *Linguamática*, 12(1):63–79, 2020.

[45] Cristiano R. de Carvalho, Edleno Silva de Moura, Adriano Veloso, and Nivio Ziviani. Website replica detection with distant supervision. *Inf. Retr. J.*, 21(4):253–272, 2018.

[46] Túlio C. Loures, Pedro O. S. Vaz de Melo, and Adriano Alonso Veloso. Is it possible to describe television series from online comments? *J. Internet Serv. Appl.*, 9(1):1–17, 2018.

[47] Keiller Nogueira, Adriano Alonso Veloso, and Jefersson Alex dos Santos. Pointwise and pairwise clothing annotation: combining features from social media. *Multim. Tools Appl.*, 75(7):4083–4113, 2016.

[48] 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.

SafeML: Aprendizado de modelos preditivos seguros com restrições quanto ao mecanismo de predição

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 das predições de forma a produzir modelos preditivos seguros e/ou mais eficazes. 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 preditivos seguros, 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 um modelo preditivo vem caindo muito rapidamente. A redução do custo, por sua vez, torna cada vez mais pervasiva a utilização de modelos preditivos, 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 de modelos preditivos. A motivação para nosso projeto, portanto, é a necessidade de se produzir modelos preditivos com restrições quanto ao mecanismo de predição, 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 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.

Relação com o Projeto Anterior (Agosto 2017 – Julho 2020): O projeto anterior foi focado no desenvolvimento de algoritmos de Aprendizado Profundo utilizando novas arquiteturas aprimoradas a partir de redes recorrentes, convolucionais, adversariais, sequência-sequência, treinadas através de supervisão e adaptação de domínio, ou por reforço. Os algoritmos desenvolvidos foram avaliados em diferentes cenários de aplicação, mostrando-se especialmente efetivos em aplicações relacionadas à Visão e Linguística Computacional. Desta forma, pretendemos aprimorar nossos algoritmos incorporando restrições relacionadas à explicabilidade durante o processo de aprendizado. 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 manutenção da bolsa de produtividade em pesquisa nível 2, ou 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, após o doutorado.
- é vinculado ao Programa de Pós-Graduação do Departamento de Ciência da COmputação da UFMG desde 2011, e já orientou dissertações de mestrado e teses de doutorado.

1 Introdução

A modelagem explanatória e a modelagem preditiva são duas facetas da construção de modelos a partir de dados [71]. Embora haja uma sensação instintiva de que predizer e explicar sejam tarefas diferentes, muitas vezes presume-se que modelos com alto poder explicativo são inerentemente de alto poder preditivo. Ainda assim, a maior parte da literatura recente não explora nenhum tipo de relação entre explicação e predição durante a construção de modelos a partir dos dados.

Pelo contrário, os avanços na área de Aprendizado de Máquina propociaram, na verdade, uma drástica redução do custo da predição. Em outras palavras, a construção de modelos preditivos eficazes vem se tornando um processo cada vez mais conhecido e adotado. Porém, com o aumento na adoção desses modelos vieram à tona as mais diversas preocupações acerca dos mecanismos por trás das predições sendo realizadas. Como o uso de modelos preditivos muitas vezes visa a automatização de processos, mecanismos de predição que sejam injustos ou incorretos irão fatalmente aumentar a escala na qual problemas como discriminação, intolerância e desigualdade afetam nossa sociedade.

Em resposta à necessidade de verificação e entendimento dos mecanismos por trás de um modelo preditivo, uma série de diferentes modelos explanatórios foram propostos recentemente. Tais modelos explanatórios geralmente estimam a importância e o efeito das características sendo empregadas pelo modelo preditivo [57, 60, 68, 69]. Dessa forma, modelos preditivos podem ser avaliados *a posteriori* sob a luz de seu respectivo mecanismo de predição. Por exemplo, pode-se estimar como diferentes fatores afetam o avanço da pandemia do novo coronavírus [75], ou quais marcadores periféricos estão associados ao diagnóstico da Doença de Alzheimer [63]. Contudo, observamos que os modelos explanatórios vêm sendo utilizados exclusivamente para verificar os mecanismos de predição, e dessa forma, uma vez constatadas inconsistências com o mecanismo de predição, não há nenhum tipo de auxílio no sentido de como proceder de maneira a aprimorar ou até mesmo corrigir o modelo.

É nesse sentido que acreditamos existir uma relação mais virtuosa entre modelos preditivos e explanatórios, e que podemos tirar proveito de tal relação de forma a produzir modelos mais confiáveis, justos e interpretáveis. 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 empregada pelo modelo preditivo. Desta forma, assumimos a disponibilidade de uma visão (mesmo que parcial) em relação às explicações esperadas (ou corretas) [54], 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

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

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 preditivos 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.

- 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.

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. 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 preditivos 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 das predições de forma a produzir modelos preditivos mais alinhados com o conhecimento prévio sobre o fenômeno sendo modelado. Atualmente, a busca por modelos preditivos em geral dá-se unicamente em termos da tentativa de escolher o modelo que forneça a menor expectativa de erro de predição. No entanto, a comunidade vem observando de maneira crescente, que essa abordagem pode resultar em modelos eficazes, porém injustos ou incorretos [52, 53, 55, 56, 62, 64]. A literatura mais recente já contempla iniciativas demonstrando a possibilidade de se construir modelos preditivos a partir de explicações [58]. 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 sejam considerados seguros.

2.1 Restrições quanto ao mecanismo de predição

A base de nosso projeto é a possibilidade de podermos associar um modelo explanatório a um modelo preditivo, de forma que o modelo explanatório revele o mecanismo de predição empregado pelo modelo preditivo. 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 preditivos. Mais especificamente, o conhecimento prévio produziria restrições e mecanismos de predição válidos precisariam estar em concordância com essas restrições. Sendo assim, dentre os muitos modelos preditivos que poderiam ser extraídos dos dados, só seriam considerados aqueles cujos mecanismos de predição 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

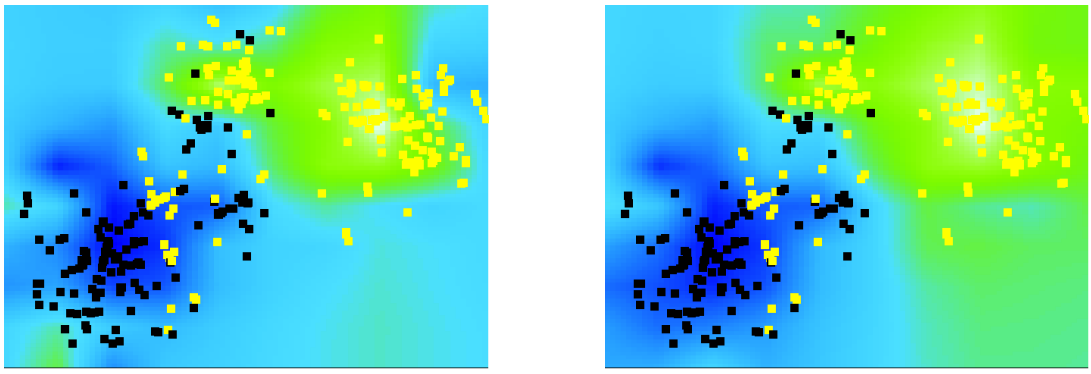

Figura 1: Separações produzidas por dois modelos preditivos diferentes. Na esquerda o modelo preditivo foi obtido sem uso de restrições. Na direita o modelo preditivo foi obtido empregando-se restrições nas características "cortisol" e "proteína C-reativa".

avançaram. Os dois modelos apresentam erros de predição em termos de área abaixo da curva ROC [51] muito similares, porém produzem separações ligeiramente diferentes, como mostrado na figura. A Figura 2 mostra o modelo explanatório [57] que foi associado ao modelo preditivo selecionado.

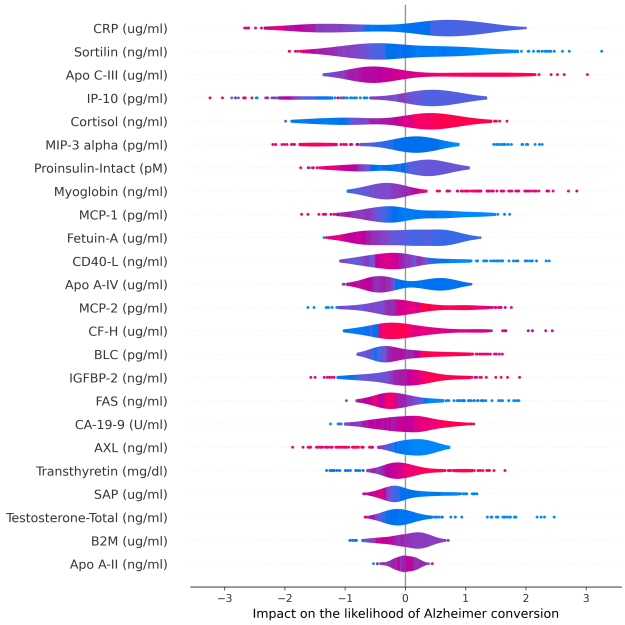

Figura 2: Mecanismo de predição empregado pelo modelo obtido empregando-se restrições nas características "cortisol" e "proteína C-reativa".

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. Certamente, 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 preditivos

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 [59]. Dessa forma, um modelo preditivo 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 preditivos pode ser explorado variando-se o valor dos hiperparâmetros, bem como selecionando as características desejadas para compor o modelo [49]. A busca no espaço de modelos preditivos geralmente ocorre exclusivamente através da minimização do erro de predição esperado, resultando assim em modelos com aparente baixo erro esperado. Ao empregar restrições durante o processo de busca por modelos preditivos, estamos diminuindo o número de modelos a serem considerados. Basicamente, o objetivo continua sendo encontrar o modelo preditivo com o menor erro esperado, porém dentro do espaço de modelos com mecanismos de predição considerados válidos dado o conjunto de restrições. Mais especificamente, a toda vez que um modelo preditivo candidato for obtido, será construído também um modelo explanatório de forma a evidenciar o mecanismo de predição do modelo preditivo. Dessa forma, as restrições podem ser comparadas ao mecanismo de predição, validando ou invalidando o modelo preditivo.

Dependendo da complexidade do fenômeno sendo modelado, é comum que sejam encontrados modelos preditivos 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 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 do modelo preditivo. Pretendemos assim, comparar modelos preditivos 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

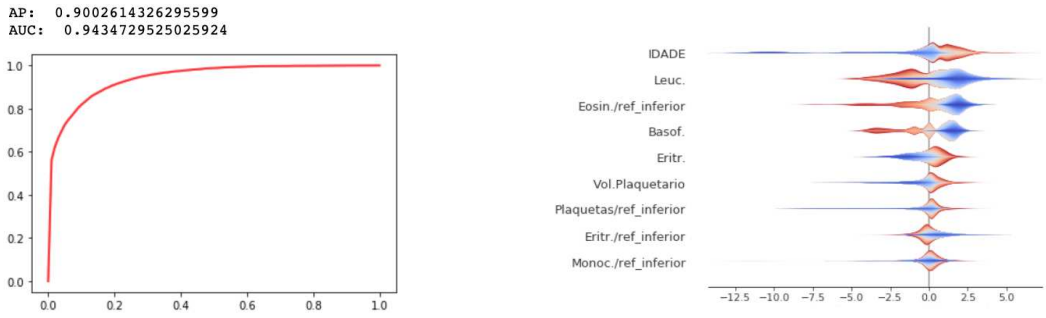

Figura 3: À esquerda, o desempenho esperado para o modelo. À direita, o mecanismo de predição do modelo.

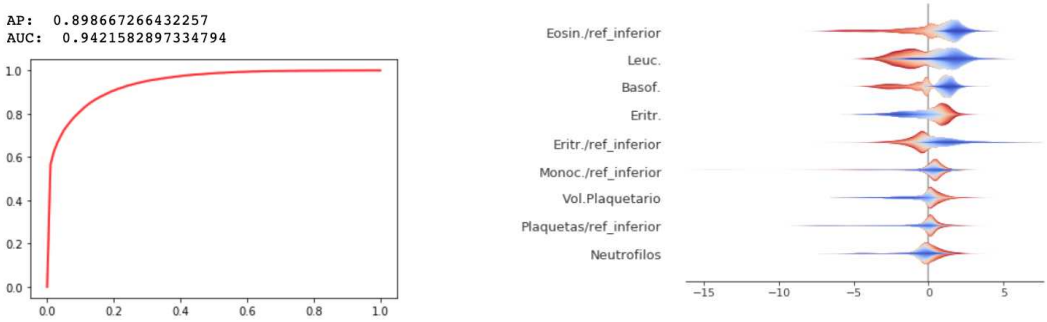

Figura 4: À esquerda, o desempenho esperado para o modelo. À direita, o mecanismo de predição do modelo.

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. Sendo assim, dizemos que uma característica tem efeito negativo na predição se ela contribui para dimunir o valor 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 [50] 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 [74] para a avaliação e comparação da qualidade das explicações fornecidas por algoritmos como Smooth Grad [72], Layer-Wise Relevance Propagation [60], Gradient Propagation [73] e Grad-CAM [70]. Sendo assim, dada nossa experiência em construir modelos preditivos e explanatórios, é natural propor a exploração de possíveis relações entre esses dois tipos de modelagem de forma a obter modelos preditivos seguros, confiáveis e eficazes.

Metodologia. Para o desenvolvimento de nossos algoritmos utilizaremos os pacotes shap, scikit-learn, e Pytorch. 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 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 preditivos com restrições quanto ao mecanismo de predição empregado. 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 preditivos com restrições de explicação. Neste caso, as restrições devem ser previamente estabelecidas, e um modelo é válido apenas se os mecanismos por trás das predições respeitem tais restrições.
- 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

Este projeto tem por objetivo principal a progressão para a categoria 1D, ou a renovação da bolsa de produtividade em pesquisa do proponente, 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.

Referências

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

[50] 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.

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

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

[60] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209. 2019.

[61] Mariane Moreira, Jefersson Alex dos Santos, and Adriano Veloso. Learning to rank similar apparel styles with economically-efficient rule-based active learning. In *International Conference on Multimedia Retrieval, ICMR*, page 361, 2014.

[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 alzheimer's 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 alzheimer's 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] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV*, pages 618–626, 2017.

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

[72] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[73] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR*, 2015.

[74] 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.

[75] Adriano Veloso and Nivio Ziviani. Explainable death toll motion modeling: Covid-19 narratives and counterfactuals. *medRxiv*, 2020.